How to compare different loss functions and their risks

Ingo Steinwart
Modeling, Algorithms and Informatics Group, CCS-3
Los Alamos National Laboratory
{ingo}@lanl.gov

September 7, 2006

Abstract

Many learning problems are described by a risk functional which in turn is defined by a loss function, and a straightforward and widely-known approach to learn such problems is to minimize a (modified) empirical version of this risk functional. However, in many cases this approach suffers from substantial problems such as computational requirements in classification or robustness concerns in regression. In order to resolve these issues many successful learning algorithms try to minimize a (modified) empirical risk of a surrogate loss function, instead. Of course, such a surrogate loss must be "reasonably related" to the original loss function since otherwise this approach cannot work well. For classification good surrogate loss functions have been recently identified, and the relationship between the excess classification risk and the excess risk of these surrogate loss functions has been exactly described. However, beyond the classification problem little is known on good surrogate loss functions up to now. In this work we establish a general theory that provides powerful tools for comparing excess risks of different loss functions. We then apply this theory to several learning problems including (cost-sensitive) classification, regression, density estimation, and density level detection.

1 Introduction

In many machine learning problems the learning goal is described by a loss function and its associated risk. A typical example of such a learning problem is *binary classification*, where a training set $T := ((x_1, y_1), \ldots, (x_n, y_n))$ is given and the goal is to predict the label $y \in Y := \{-1, 1\}$ for a new, unseen input sample $x \in X$. Commonly, it is assumed that the samples are drawn in an i.i.d. fashion from an unknown probability measure P on $X \times Y$. The simplest learning goal is then to have an almost minimal prediction error on average future data, i.e. to find a function $f : X \to \mathbb{R}$ such that

$$P\Big(\big\{(x, y) \in X \times Y : \operatorname{sign} f(x) \neq y\big\}\Big)$$

is as small as possible. It is well-known and obvious that writing $L(y, t) := 1$ if $y \operatorname{sign} t < 0$ and $L(y, t) := 0$ otherwise, the above probability equals the so-called classification risk

$$\mathcal{R}_{L,P}(f) := \int_{X \times Y} L\big(y, f(x)\big)\, dP(x, y)$$

and consequently, the learning goal is then to find a function f that (approximately) minimizes this risk. This form of a learning goal appears in many other learning problems, too. Probably the best known example of such a learning problem is real-valued *regression*, however there are many

other important learning problems like *cost-sensitive binary classification*, *multi-class classification*, *density estimation*, or *density level detection* (see Section 4 for a rigorous definition of these learning problems) which can also be described by a risk functional.

Having a learning problem of the above form a simple and well-known learning approach is the empirical risk minimization (ERM) method. Basically, the idea of ERM is to replace the unknown true risk $\mathcal{R}_{L,P}(.)$ by its empirical counterpart based on the training set T and minimize this empirical L-risk over a suitable function class. Though this approach has a lot of theoretical merits, for classification it typically leads to NP-hard optimization problems (see e.g. [14]), and thus it is not computationally realizable. One way to resolve this issue is to use a *surrogate loss function* L_2, e.g. the hinge loss as in support vector machines, and then to minimize the (perhaps even further modified) empirical L_2-risk. Let us assume for a moment that such a learning method $\mathcal{L}$ learns the surrogate learning problem defined by the loss L_2, i.e. we know $\mathcal{R}_{L_2,P}(f_T) \to \mathcal{R}^*_{L_2,P}$, where f_T is the function produced by the learning method $\mathcal{L}$ and $\mathcal{R}^*_{L_2,P} := \inf\{\mathcal{R}_{L_2,P}(f) \,|\, f : X \to \mathbb{R} \text{ measurable}\}$ is the smallest possible L_2-risk. The first question which then naturally arises is:

Question 1. *Does the convergence* $\mathcal{R}_{L_2,P}(f_T) \to \mathcal{R}^*_{L_2,P}$ *imply the convergence* $\mathcal{R}_{L,P}(f_T) \to \mathcal{R}^*_{L,P}$*?*

If we can find a positive answer to this question we know at least asymptotically that by solving the surrogate learning problem we actually also solve the learning problem we are interested in, and in this sense $\mathcal{L}$ is a suitable learning approach. However, in many situations we are not only interested in such an asymptotic relation, but also in a more quantitative statement. This leads to the second question:

Question 2. *Does there exists an increasing function* $\Upsilon : [0,\infty) \to [0,\infty)$ *with* $\Upsilon(0) = 0$*, which is continuous in 0 and satisfies*

$$\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} \;\le\; \Upsilon\big(\mathcal{R}_{L_2,P}(f) - \mathcal{R}^*_{L_2,P}\big) \;?$$

For the binary classification problem both questions have been intensively investigated in recent years (see e.g. [18, 19, 37, 27, 1]), and for margin-based loss functions L_2, i.e. for loss functions of the form $L_2(y,t) = \varphi(yt)$, complete answers were established in [1]. However, using a surrogate loss function is a strategy that is not only interesting for binary classification. Indeed, this approach also has its benefits in the following learning problems:

Cost-sensitive binary classification. In cost-sensitive binary classification the two different types of errors are penalized differently. Consequently, the optimization problem of a simple ERM approach remains NP-hard to solve, and thus suitable surrogate loss functions are desired for a computational treatment. Despite the practical importance of this learning problem only very little is known in view of the above questions (see [20] for a conjecture related to the above questions and [10] for a discussion of this learning problem).

Regression. The classical loss function for regression problems is the least squares loss. Though this loss is mathematically rather easy to handle and the corresponding learning algorithms are often computational feasible, it is well-known that minimizing an empirical risk based on the least squares loss is a method which is quite sensitive to outliers. Therefore, a number of more robust surrogate loss functions have been proposed in the last decades (see e.g. [15, 32, 25, 9]). Up to now, the analysis of such regression loss functions is typically conducted in a maximum-likelihood fashion which relates the loss functions to density models of the (conditional) noise distributions. However, the maximum-likelihood approach requires knowledge about

the marginal distribution which is usually not available and in addition, it does not give information regarding our two basic questions.

Density level detection. The common performance measure of the density level detection problem is a risk functional whose loss function is defined in terms of the *unknown* density (see e.g. [12, 21, 22, 31]). Consequently, the loss function is unknown and cannot be used to *a)* compute a test error, *b)* build a (modified) ERM approach, and *c)* compare different solutions to the density level detection problem. Therefore, having a reasonable surrogate risk is essential for dealing with the density level detection problem.

Density estimation. The common performance measures for the density estimation problem are p-norms of the difference between the unknown density and its estimate. Again, these performance measures are risks with respect to a loss function defined by the unknown density, and consequently, the density estimation problem suffers from the same problems as the density level detection problem does.

Obviously, the above learning problems are rather different regarding their defining risk functionals, and in addition, the above list is by no means complete (see e.g. the recent work on multi-class classification in [36, 30]). In order to systematically investigate our two questions we will therefore first establish a very general theory that describes how to relate the excess risks of different loss functions. In large parts this general theory unifies earlier findings of [1], [27], [37], [36], but it also contains new results. In a second step we then demonstrate the power of this theory by applying it to the learning problems described above. Let us briefly summarize our findings:

- **Cost-sensitive classification.** We show that a natural weighting method for margin-based loss functions allows us to translate the results of [1] to analogous results for cost-sensitive classification. In particular, we give positive answers to Question 1 and 2. Moreover, we show that the natural weighting method is the only one that allows such positive answers.

- **Regression.** We first show that the least squares loss is essentially the only loss function of the form $L_2(y,t) = \psi(y-t)$ that can be used to find the regression function. For some large classes of *symmetric* noise distributions we then characterize the loss functions of the above form that allow positive answers to the questions 1 and 2 if the target risk is the 1-norm distance between the regression function and its prediction. Here it will turn out that the *convexity* of L_2 and related, stronger notions such as strict convexity and uniform convexity play a crucial role. Moreover, we show that for unbounded, symmetric, but otherwise unspecified noise every loss function of the above form that admits a positive answer to Question 2 must grow at least as fast as the squared loss. Consequently, every ERM approach based on such loss functions for finding the regression function either learns the problem only in a weak sense or is sensitive to outliers. Finally, we discuss in which cases and in which sense approximate risk minimizers approximate the exact risk minimizer.

- **Density level detection.** It was recently shown in [28] that the density level detection problem can be solved by learning a binary classification problem, and in fact, the widely known excess mass approach (see e.g. [12, 21, 22, 31]) implicitly implements ERM of this classification problem. We extend the considerations of [28] and show that there exists no surrogate loss function which allows a positive answer to the second question without making assumptions on the density.

- **Density estimation.** A well-known heuristic (see e.g. [13, Chap. 14.2.4]) casts the density estimation problem into a supervised learning problem by using additional samples drawn

from the known reference measure. However, to our best knowledge, there is nothing known about this heuristic in view of the above questions. For convex loss functions we are able to give a weak positive answer to the first question. Furthermore, we present a result showing that the second answer cannot be positively answered.

The rest of this work is organized as follows: In the sections 2 and 3 we develop the general theory on excess risks. Section 4 then contains the results for the above learning problems. The proofs of the results of the sections 2 and 3 can be found in Section 5. In addition, this section contains a powerful tool ensuring the existence of measurable selections which is used to deal with the notoriously unpleasant measurability questions related to minimal risks. Finally, the appendix contains some additional information on stronger notions of convexity.

2 The General Theory

In this section we introduce some fundamental definitions related to general cost functions of a form similar to those considered in [33]. We then develop the central tools for investigating surrogate cost functions and present general answers to Questions 1 and 2. Some of these answers are not genuinely new and have been previously found in other forms or less generality by other authors (see e.g. [1], [27], [37], and [36]), and the proofs of these results are not that deep, either. What is new, however, is the language we introduce in this section which clearly describes the key quantity related to surrogate cost functions, namely the calibration function defined in Lemma 2.9. In particular, this calibration function will enable us to easily investigate the general case in this section and the following section, and important examples in Section 4. Consequently, it is *the* central notion of the entire work.

Let us begin with some notations which will be used throughout this work. To this end let X be a nonempty set equipped with a σ-algebra $\mathcal{X}$. Given a finite measure μ on $\mathcal{X}$ the μ-completion $\mathcal{X}_\mu$ of $\mathcal{X}$ is defined by

$$\mathcal{X}_\mu := \left\{A \cup B : A \in \mathcal{X}, \exists N \in \mathcal{X} \text{ with } \mu(N) = 0 \text{ and } B \subset N\right\}.$$

Moreover, the completion $\hat{\mathcal{X}}$ of $\mathcal{X}$ is defined by

$$\hat{\mathcal{X}} := \left\{A \subset X : A \in \mathcal{X}_\mu \text{ for all finite measures } \mu \text{ on } \mathcal{X}\right\},$$

and the measurable space $(X, \mathcal{X})$ is called *complete* if $\mathcal{X} = \hat{\mathcal{X}}$. Throughout this work we assume that $(X, \mathcal{X})$ is complete. Moreover, Y always denotes Polish space, i.e. a topological space whose topology can be described by a complete and separable metric, and we always equip Y with its Borel σ-algebra. In particular recall that all open or closed subsets of $\mathbb{R}$ are Polish spaces.

For a probability measure P on $X \times Y$ we denote the marginal distribution on X by P_X. Furthermore, $P(\,.\,|\,.\,) : \mathcal{Y} \times X \to [0,1]$ stands for a *fixed* regular conditional probability, so that we have

$$\int_{X\times Y} f\, dP = \int_X \int_Y f(x,y)\, dP(y|x) dP_X(x)$$

for all measurable functions $f : X \times Y \to [0,\infty]$. As usual $\mathcal{L}_p(\mu)$, $p \in (0,\infty)$, denotes the space of all measurable functions $f : X \to \mathbb{R}$ that are p-integrable with respect to the measure μ on $(X, \mathcal{X})$. Finally, throughout this section $\mathcal{A}$ denotes a non-empty but otherwise arbitrary set which describes the parameter or function which we wish to estimate. Let us now begin with some fundamental definitions:

Definition 2.1 (Cost function) *A function $L : X \times Y \times \mathcal{A} \to [0, \infty]$ is called a* cost function *if $L(., ., \alpha) : X \times Y \to [0, \infty]$ is measurable for all $\alpha \in \mathcal{A}$*

Cost functions which play a key role in most machine learning considerations measure the cost of using the parameter α at the point (x, y). In general we are interested in small *average* costs which are introduced in the following definition:

Definition 2.2 (Risk) *Let $L : X \times Y \times \mathcal{A} \to [0, \infty]$ be a cost function and P be a distribution on $X \times Y$. We define the L-*risk *of $\alpha \in \mathcal{A}$ by*

$$\mathcal{R}_{L,P}(\alpha) := \int_{X \times Y} L(x, y, \alpha) dP(x, y) = \int_X \int_Y L(x, y, \alpha) dP(y|x) dP_X(x) \,. \tag{1}$$

Moreover, the minimal L-risk*, also called the* Bayes L-risk*, is denoted by $\mathcal{R}^*_{L,P} := \inf_{\alpha \in \mathcal{A}} \mathcal{R}_{L,P}(\alpha)$.*

The risk $\mathcal{R}_{L,P}(\alpha)$ obviously describes the average cost of using the parameter α, where the average is taken with respect to the distribution P. Now note that using the regular conditional probability $P(y|x)$, the risk $\mathcal{R}_{L,P}(\alpha)$ can be computed by an iterated integral as we have seen in (1). Since the inner integral in (1) will be of fundamental importance in our analysis we introduce:

Definition 2.3 (Inner risk) *Let $L : X \times Y \times \mathcal{A} \to [0, \infty]$ be a cost function and Q be a distribution on Y. We define the* inner L-risk *of an element $\alpha \in \mathcal{A}$ by*

$$\mathcal{C}_{L,Q,x}(\alpha) := \int_Y L(x, y, \alpha) dQ(y) \qquad\qquad x \in X.$$

Furthermore, the minimal inner L-risk *is denoted by $\mathcal{C}^*_{L,Q,x} := \inf_{\alpha \in \mathcal{A}} \mathcal{C}_{L,Q,x}(\alpha)$.*

Note that given a distribution P on $X \times Y$ the inner risks $\mathcal{C}_{L,P(.|x),x}(\alpha)$, $x \in X$, of α can be used to compute the risk $\mathcal{R}_{L,P}(\alpha)$ since (1) immediately gives

$$\mathcal{R}_{L,P}(\alpha) = \int_X \mathcal{C}_{L,P(.|x),x}(\alpha) \, dP_X(x) \,.$$

Our first goal is to establish the same relation between the minimal inner risks and the minimal risk. To this end we need the following definition which will be used to ensure that $x \mapsto \mathcal{C}^*_{L,P(.|x),x}$ is measurable:

Definition 2.4 (Minimizable cost functions) *Let $L : X \times Y \times \mathcal{A} \to [0, \infty]$ be a cost function and P be a distribution on $X \times Y$. We say that L* is P-minimizable *if for all $\varepsilon > 0$ there exists an $\alpha_\varepsilon \in \mathcal{A}$ such that for all $x \in X$ we have*

$$\mathcal{C}_{L,P(.|x),x}(\alpha_\varepsilon) < \mathcal{C}^*_{L,P(.|x),x} + \varepsilon \,. \tag{2}$$

Note that the above definition in particular ensures $\mathcal{C}^*_{L,P(.|x),x} < \infty$ for all $x \in X$. Now we can establish the following simple but important lemma which computes $\mathcal{R}^*_{L,P}$ with the help of the corresponding minimal inner risks:

Lemma 2.5 *Let P be a distribution on $X \times Y$ and $L : X \times Y \times \mathcal{A} \to [0, \infty]$ be a P-minimizable cost function. Then $x \mapsto \mathcal{C}^*_{L,P(.|x),x}$ is measurable and we have*

$$\mathcal{R}^*_{L,P} = \int_X \mathcal{C}^*_{L,P(.|x),x} \, dP_X(x) \,.$$

The above lemma shows that the minimal risk $\mathcal{R}^*_{L,P}$ can be achieved by *pointwisely minimizing* the inner risks $\mathcal{C}_{L,P(.|x),x}(.)$, $x \in X$, which—in general—will be easier than direct minimization of $\mathcal{R}_{L,P}(.)$. Moreover, for P-minimizable L with $\mathcal{R}^*_{L,P} < \infty$ we have

$$\mathcal{R}_{L,P}(\alpha) - \mathcal{R}^*_{L,P} = \int_X \mathcal{C}_{L,P(.|x),x}(\alpha) - \mathcal{C}^*_{L,P(.|x),x}\, dP_X(x)$$

for all $\alpha \in \mathcal{A}$. Consequently, we can split the analysis of $\mathcal{R}_{L,P}(\alpha) - \mathcal{R}^*_{L,P}$ into *a)* the analysis of the inner excess risks $\mathcal{C}_{L,P(.|x),x}(\alpha) - \mathcal{C}^*_{L,P(.|x),x}$, $x \in X$, and *b)* the investigation of the integration with respect to P_X. Besides technical benefits, the major benefit of this approach is that the analysis in *a)* only depends on P via the conditional distributions $P(.|x)$ and hence we can consider the excess inner risks $\mathcal{C}_{L,Q,x}(\alpha) - \mathcal{C}^*_{L,Q,x}$ for suitable distributions Q on Y as a template for $\mathcal{C}_{L,P(.|x),x}(\alpha) - \mathcal{C}^*_{L,P(.|x),x}$. The latter does not only reduce notations but also supports the machine learning point of view in which it is assumed that P, and hence $P(.|x)$, $x \in X$, is (almost) completely unknown. To pursue this idea we make the following definition:

Definition 2.6 *Let $\mathcal{Q}$ be a set of distributions on Y. We say that a distribution P on $X \times Y$ is* of type $\mathcal{Q}$ *if $P(.|x) \in \mathcal{Q}$ for all $x \in X$.*

In many machine learning problems the only information given about the distribution P is that it is of type $\mathcal{Q}$ for some "large" set of distributions $\mathcal{Q}$. For example, in binary classification one typically only knows that P is a distribution on $X \times \{-1, 1\}$, i.e. that P is a distribution of type $\mathcal{Q}$, where $\mathcal{Q}$ consists of all distributions on $\{-1, 1\}$.

Before we begin with our analysis let us introduce some more notations which will be very useful. The sets containing the elements in $\mathcal{A}$ that "almost" minimize the inner risk at x are denoted by

$$\mathcal{M}_{L,Q,x}(\varepsilon) := \left\{\alpha \in \mathcal{A} : \mathcal{C}_{L,Q,x}(\alpha) < \mathcal{C}^*_{L,Q,x} + \varepsilon\right\}, \qquad \varepsilon \in [0, \infty].$$

For later use we note that we always have $\mathcal{M}_{L,Q,x}(0) = \emptyset$ and $\mathcal{M}_{L,Q,x}(\varepsilon_1) \subset \mathcal{M}_{L,Q,x}(\varepsilon_2)$ for $0 \leq \varepsilon_1 \leq \varepsilon_2 \leq \infty$. Furthermore we have $\mathcal{M}_{L,Q,x}(\varepsilon) \neq \emptyset$ for some $\varepsilon \in (0, \infty]$ if and only if $\mathcal{C}^*_{L,Q,x} < \infty$. In particular, for P-minimizable cost functions we have $\mathcal{M}_{L,P(.|x),x}(\varepsilon) \neq \emptyset$ for all $\varepsilon > 0$ and $x \in X$. Finally, we write

$$\mathcal{M}_{L,Q,x}(0^+) := \bigcap_{\varepsilon > 0} \mathcal{M}_{L,Q,x}(\varepsilon) = \left\{\alpha \in \mathcal{A} : \mathcal{C}_{L,Q,x}(\alpha) < \mathcal{C}^*_{L,Q,x} + \varepsilon \text{ for all } \varepsilon > 0\right\}.$$

Note that in the case $\mathcal{C}^*_{L,Q,x} < \infty$ the set $\mathcal{M}_{L,Q,x}(0^+)$ contains those elements in $\mathcal{A}$ that *exactly* minimize $\mathcal{C}_{L,Q,x}(.)$, while in the other case $\mathcal{C}^*_{L,Q,x} = \infty$ the set is empty.

Let us now turn to the main goal of this section, namely comparing the excess risks of different cost functions. To this end we first observe that given two cost functions $L_1 : X \times Y \times \mathcal{A} \to [0, \infty)$ and $L_2 : X \times Y \times \mathcal{A} \to [0, \infty)$ with

$$\mathcal{M}_{L_2,P(.|x),x}(0^+) \subset \mathcal{M}_{L_1,P(.|x),x}(0^+), \qquad x \in X,$$

Lemma 2.5 shows that every exact minimizer $\alpha^* \in \mathcal{A}$ of $\mathcal{R}_{L_2,P}(.)$ is also an exact minimizer of $\mathcal{R}_{L_1,P}(.)$. However, exact minimizers do not necessarily exist, and even if they do exist it is rather unlikely that we will find them by a learning procedure. On the other hand, many learning procedures guarantee to find approximate minimizers with high probability, and thus our overall goal is to establish properties similar to the above observation for *approximate* minimizers. Let us begin with the following definition.

Definition 2.7 (Calibration) *Let $\mathcal{Q}$ be a set of distributions on Y, and $L_1 : X \times Y \times \mathcal{A} \to [0,\infty]$, $L_2 : X \times Y \times \mathcal{A} \to [0,\infty]$ be two cost functions. We say that L_2* is L_1-calibrated with respect to $\mathcal{Q}$ *if for all $\varepsilon > 0$, $Q \in \mathcal{Q}$, and $x \in X$ there exists a $\delta > 0$ such that*

$$\mathcal{M}_{L_2,Q,x}(\delta) \subset \mathcal{M}_{L_1,Q,x}(\varepsilon)\,, \tag{3}$$

i.e. if for all $\alpha \in \mathcal{A}$ we have

$$\mathcal{C}_{L_2,Q,x}(\alpha) < \mathcal{C}^*_{L_2,Q,x} + \delta \qquad \Longrightarrow \qquad \mathcal{C}_{L_1,Q,x}(\alpha) < \mathcal{C}^*_{L_1,Q,x} + \varepsilon\,. \tag{4}$$

Furthermore, we say that L_2 is L_1-calibrated with respect to a distribution P on $X \times Y$ if L_2 is L_1-calibrated with respect to the set $\{P(.|x) : x \in X\}$.

Obviously, L_2 is L_1-calibrated with respect to P if for any given accuracy $\varepsilon > 0$ there exists a $\delta > 0$ such that every $\alpha \in \mathcal{A}$ minimizing $\mathcal{C}_{L_2,P(.|x),x}(.)$ up to δ minimizes $\mathcal{C}_{L_1,P(.|x),x}(.)$ up to the desired accuracy ε. In other words, L_2 is L_1-calibrated with respect to P if and only if we have a positive answer to Question 1 for the excess *inner risks*. Before we investigate some general techniques to check for calibration we now present our first main result that shows that for calibrated cost functions we often have a positive answer to Question 1 for the excess *risks*, too:

Theorem 2.8 *Let P be a distribution on $X \times Y$ and $L_1 : X \times Y \times \mathcal{A} \to [0,\infty]$, $L_2 : X \times Y \times \mathcal{A} \to [0,\infty]$ be two P-minimizable cost functions such that $\mathcal{R}^*_{L_1,P} < \infty$ and $\mathcal{R}^*_{L_2,P} < \infty$. Furthermore assume that there exist a function $b \in \mathcal{L}_1(P_X)$ and measurable functions $\delta(\varepsilon,.) : X \to (0,\infty)$, $\varepsilon > 0$, such that*

$$\mathcal{C}_{L_1,P(.|x),x}(\alpha) \;\le\; \mathcal{C}^*_{L_1,P(.|x),x} + b(x) \tag{5}$$

and

$$\mathcal{C}_{L_2,P(.|x),x}(\alpha) < \mathcal{C}^*_{L_2,P(.|x),x} + \delta(\varepsilon,x) \qquad \Longrightarrow \qquad \mathcal{C}_{L_1,P(.|x),x}(\alpha) < \mathcal{C}^*_{L_1,P(.|x),x} + \varepsilon \tag{6}$$

for all $x \in X$, $\varepsilon > 0$ and $\alpha \in \mathcal{A}$. Then for all $\varepsilon > 0$ there exists a $\delta > 0$ such that for all $\alpha \in \mathcal{A}$ we have

$$\mathcal{R}_{L_2,P}(\alpha) < \mathcal{R}^*_{L_2,P} + \delta \qquad \Longrightarrow \qquad \mathcal{R}_{L_1,P}(\alpha) < \mathcal{R}^*_{L_1,P} + \varepsilon\,. \tag{7}$$

Note that (7) is equivalent to a positive answer of Question 1, and consequently Theorem 2.8 gives a sufficient condition for a asymptotic relationship between different excess risks in the sense of this question. Moreover, Condition (6) implies that L_2 is L_1-calibrated with respect to P, and in fact, the only difference between this condition and the calibration with respect to P is the *measurability* of $\delta(\varepsilon,.)$ which is necessary for technical reasons. However, we will see later in Theorem 3.3 that for $\mathcal{A}$ consisting of *functions over* X, this measurability is often automatically satisfied by the so-called *calibration function* defined below. In addition, Theorem 3.3 shows that in this case the L_1-calibration of L_2 is also a *necessary* condition for (7), and hence this theorem will give us a complete answer to Question 1.

In order to apply Theorem 2.8 the main difficulty is usually to determine numbers $\delta(\varepsilon,x) > 0$ with $\mathcal{M}_{L_2,P(.|x),x}(\delta(\varepsilon,x)) \subset \mathcal{M}_{L_1,P(.|x),x}(\varepsilon)$, i.e. to ensure the calibration. The following lemma describes an easy yet optimal solution for this task.

Lemma 2.9 (Calibration function) *Let Q be a distribution on Y, and $L_1 : X \times Y \times \mathcal{A} \to [0,\infty]$, $L_2 : X \times Y \times \mathcal{A} \to [0,\infty]$ be two cost functions. Then we define the* calibration function $\delta_{\max}(.,Q,x) : [0,\infty] \to [0,\infty]$ of (L_1,L_2) *by*

$$\delta_{\max}(\varepsilon,Q,x) := \begin{cases} \inf_{\alpha \in \mathcal{A} \setminus \mathcal{M}_{L_1,Q,x}(\varepsilon)} \mathcal{C}_{L_2,Q,x}(\alpha) - \mathcal{C}^*_{L_2,Q,x} & \text{if } \mathcal{C}^*_{L_2,Q,x} < \infty, \\ \infty & \text{if } \mathcal{C}^*_{L_2,Q,x} = \infty \end{cases} \tag{8}$$

for all $\varepsilon \in [0,\infty]$. Then for all $\varepsilon \in [0,\infty]$ we have:

i) $\mathcal{M}_{L_2,Q,x}(\delta_{\max}(\varepsilon,Q,x)) \subset \mathcal{M}_{L_1,Q,x}(\varepsilon)$.

ii) $\mathcal{M}_{L_2,Q,x}(\delta) \not\subset \mathcal{M}_{L_1,Q,x}(\varepsilon)$ *whenever* $\delta > \delta_{\max}(\varepsilon,Q,x)$.

In addition, if both $\mathcal{C}^*_{L_1,Q,x} < \infty$ and $\mathcal{C}^*_{L_2,Q,x} < \infty$, then for all $\alpha \in \mathcal{A}$ we have

$$\delta_{\max}\left(\mathcal{C}_{L_1,Q,x}(\alpha) - \mathcal{C}^*_{L_1,Q,x}, Q, x\right) \;\le\; \mathcal{C}_{L_2,Q,x}(\alpha) - \mathcal{C}^*_{L_2,Q,x}\,. \tag{9}$$

Note that in some cases we have to distinguish between different calibration functions and hence we occasionally use the notation $\delta_{\max,L_1,L_2}(.,Q,x) := \delta_{\max}(.,Q,x)$ for the calibration function $\delta_{\max}(.,Q,x)$ of (L_1,L_2).

Obviously, part *i)* of Lemma 2.9 states that L_2 is L_1-calibrated with respect to $\mathcal{Q}$ if $\delta_{\max}(\varepsilon,Q,x) > 0$ for all $x \in X$, $Q \in \mathcal{Q}$, and $\varepsilon > 0$. Moreover, part *ii)* shows there is no real number δ larger than $\delta_{\max}(\varepsilon,Q,x)$ satisfying the calibration condition (3), and consequently, L_2 is L_1-calibrated with respect to $\mathcal{Q}$ if and only if we have $\delta_{\max}(\varepsilon,Q,x) > 0$ for all $x \in X$, $Q \in \mathcal{Q}$, and $\varepsilon > 0$. In other words, the calibration function is *the* quantity we have to investigate when we want to check whether L_2 is L_1-calibrated or not (see e.g. the theorems 4.18, 4.19, and 4.29 for results in this direction).

Fortunately, it turns out that in most practical situations the computation of the calibration function is a straightforward exercise, and for the learning scenarios mentioned in the introduction the corresponding results can be found in the lemmas 4.1, 4.6, 4.16, and 4.28. Moreover, if the surrogate cost function is *convex*[1] in the sense of the following definition this computation is rather easy even in the general case as we will see below:

Definition 2.10 *A cost function $L : X \times Y \times \mathbb{R} \to [0,\infty]$ is called* (strictly) convex *if $L(x,y,.) : \mathbb{R} \to [0,\infty]$ is (strictly) convex for all $x \in X$ and $y \in Y$.*

As already indicated, the calibration function for convex surrogates can be easily computed. This is stated in the following result:

Lemma 2.11 *Let $\mathcal{Q}$ be a distribution on Y and $L_1 : X \times Y \times \mathbb{R} \to [0,\infty]$ be a cost function such that $\mathcal{M}_{L_1,Q,x}(\varepsilon)$ is a non-empty interval for some $x \in X$ and $\varepsilon > 0$. Moreover, let $L_2 : X \times Y \times \mathbb{R} \to [0,\infty]$ be a convex cost function such that $\mathcal{C}_{L_2,Q,x}(t) < \infty$ for all $t \in \mathbb{R}$. If $\mathcal{M}_{L_2,Q,x}(0^+) \subset \mathcal{M}_{L_1,Q,x}(0^+)$ then we have*

$$\delta_{\max}(\varepsilon,Q,x) = \min\Big\{\mathcal{C}_{L_2,Q,x}(\sup \mathcal{M}_{L_1,Q,x}(\varepsilon)), \mathcal{C}_{L_2,Q,x}(\inf \mathcal{M}_{L_1,Q,x}(\varepsilon))\Big\} - \mathcal{C}^*_{L_2,Q,x}\,, \tag{10}$$

where $\mathcal{C}_{L_2,Q,x}(\pm\infty) := \infty$.
In particular, if $\sup \mathcal{M}_{L_1,Q,x}(\varepsilon) \notin \mathcal{M}_{L_2,Q,x}(0^+)$ and $\inf \mathcal{M}_{L_1,Q,x}(\varepsilon) \notin \mathcal{M}_{L_2,Q,x}(0^+)$ then $\delta_{\max}(\varepsilon,Q,x) > 0$.

In the machine learning literature a cost function L_2 is often considered to be a suitable surrogate for L_1 if the relation $\mathcal{M}_{L_2,Q,x}(0^+) \subset \mathcal{M}_{L_1,Q,x}(0^+)$ holds. Of course, in general this relation is not sufficient for calibration. However, for many target cost functions the sets $\mathcal{M}_{L_1,Q,x}(\varepsilon)$ are intervals whose endpoints differ from the endpoints of $\mathcal{M}_{L_1,Q,x}(0^+)$. In addition, one is often only interested in *convex* surrogates because of algorithmic issues. Now note that in such cases the above lemma shows that $\mathcal{M}_{L_2,Q,x}(0^+) \subset \mathcal{M}_{L_1,Q,x}(0^+)$ is sufficient for calibration, and hence it gives the first rigorous justification for considering this inclusion instead of general calibration question.

[1] Basic definitions and properties regarding convexity can be found in the appendix.

Let us now investigate Question 2. To this end observe that in some sense inequalities between the involved excess risks are readily available by Theorem 2.8: indeed, for $\varepsilon > 0$ let $\delta(\varepsilon) > 0$ be a real number such that implication (7) holds for all $\alpha \in \mathcal{A}$. Furthermore, we define $\delta(0) := 0$. For $\alpha \in \mathcal{A}$ with $\varepsilon := \mathcal{R}_{L_1,P}(\alpha) - \mathcal{R}^*_{L_1,P} > 0$ we then have $\mathcal{R}_{L_2,P}(\alpha) - \mathcal{R}^*_{L_2,P} \geq \delta(\varepsilon)$, or in other words

$$\delta\big(\mathcal{R}_{L_1,P}(\alpha) - \mathcal{R}^*_{L_1,P}\big) \leq \mathcal{R}_{L_2,P}(\alpha) - \mathcal{R}^*_{L_2,P}\,. \tag{11}$$

Furthermore, for $\alpha \in \mathcal{A}$ with $\mathcal{R}_{L_1,P}(\alpha) - \mathcal{R}^*_{L_1,P} = 0$ this inequality is satisfied by our definition $\delta(0) = 0$, and consequently (11) is satisfied for all $\alpha \in \mathcal{A}$. Inverting this inequality then gives a positive answer to Question 2. However, the proof of Theorem 2.8 does unfortunately not provide a constructive way to find a value for $\delta(\varepsilon)$, and hence our next aim is to describe conditions on L_1, L_2, and P which will allow us to easily establish inequalities in many situations. To this end we begin with the following definition.

Definition 2.12 (Fenchel-Legendre biconjugate) *Let $I \subset \mathbb{R}$ be an interval and $g : I \to [0,\infty]$ be a function. Then the Fenchel-Legendre bi-conjugate $g^{**} : I \to [0,\infty]$ of g is the largest convex function $h : I \to [0,\infty]$ satisfying $h \leq g$.*

Moreover, we use the convention $g^{**}(\infty) := \lim_{x\to\infty} g^{**}(x)$ for functions $g : [0,\infty) \to [0,\infty)$.

Note that the Fenchel-Legendre bi-conjugate which is a well-known tool in convex analysis (see e.g. [24]), is determined by

$$\operatorname{Epi} g^{**} = \overline{\operatorname{co} \operatorname{Epi} g}\,,$$

where $\operatorname{Epi} g := \{(x,y) \in I \times [0,\infty] : g(x) \leq y\}$ denotes the epigraph of g and $\operatorname{co} A$ denotes the convex hull of a set A.

Now we are prepared to establish our first inequalities between excess risks of different cost functions.

Theorem 2.13 *Let P be a distribution on $X \times Y$ and $L_1 : X \times Y \times \mathcal{A} \to [0,\infty]$, $L_2 : X \times Y \times \mathcal{A} \to [0,\infty]$ be two P-minimizable cost functions with $\mathcal{R}^*_{L_1,P} < \infty$ and $\mathcal{R}^*_{L_2,P} < \infty$. We write*

$$B_\alpha := \operatorname*{ess\,sup}_{x\in X} \mathcal{C}_{L_1,P(.|x),x}(\alpha) - \mathcal{C}^*_{L_1,P(.|x),x}$$

for all $\alpha \in \mathcal{A}$. Furthermore, let $\delta : [0,\infty] \to [0,\infty]$ be a function with $\delta(0) = 0$ and

$$\mathcal{M}_{L_2,P(.|x),x}(\delta(\varepsilon)) \subset \mathcal{M}_{L_1,P(.|x),x}(\varepsilon) \tag{12}$$

for all $x \in X$ and all $\varepsilon \in [0,\infty]$. Then for all $\alpha \in \mathcal{A}$ we have

$$\delta^{**}_{B_\alpha}\big(\mathcal{R}_{L_1,P}(\alpha) - \mathcal{R}^*_{L_1,P}\big) \leq \mathcal{R}_{L_2,P}(\alpha) - \mathcal{R}^*_{L_2,P}\,, \tag{13}$$

where $\delta^{**}_{B_\alpha} : [0,B_\alpha] \to [0,\infty]$ denotes the Fenchel-Legendre biconjugate of the restriction $\delta_{|[0,B_\alpha]}$.

Remark 2.14 It is straightforward to see from the definition of the Fenchel-Legendre bi-conjugate that (13) actually holds for all convex functions $\tilde\delta : [0,B_\alpha] \to [0,\infty]$ satisfying $\tilde\delta \leq \delta$. Moreover, if the function δ in the above theorem is increasing with $\delta(\varepsilon) > 0$ for all $\varepsilon > 0$ and we have $B_\alpha < \infty$ then Lemma A.6 shows that its bi-conjugate satisfies $\delta^{**}_{B_\alpha}(\varepsilon) > 0$ for all $\varepsilon \in (0,B_\alpha]$. Moreover, the bi-conjugate is always convex, and since $\delta^{**}_{B_\alpha}(0) = 0$ it is also *strictly* increasing and thus injective. Consequently, if $\delta^{**}_{B_\alpha}$ is finite, then it has a continuous inverse, and hence we have found a positive answer to Question 2. However, in general this is no longer true if $B_\alpha = \infty$, as e.g. the case $\delta(\varepsilon) = \sqrt{\varepsilon}$, $\varepsilon \geq 0$, shows (see also [34, Prop. A.5] for conditions guaranteeing $\delta^{**} > 0$).

We will see in Theorem 2.17 that Condition (12) is also *necessary* if one wants to have an inequality between the two excess risks that is *independent* of the specific distribution P. Of course, from a machine learning perspective this independence is a highly desired property since in general the data-generating distribution P is assumed to be unknown. Finally, we will see in Section 4 that for binary classification the inequalities established in Theorem 2.13 coincide with the inequalities found in [1]. Since their inequalities are sharp if one does not make additional assumptions on P, we see that the inequalities of Theorem 2.13 cannot be improved in general. However, note that under certain conditions on P there are substantial improvements possible, and we will present examples of such improvements after Theorem 3.9.

Using the notion of uniform calibration introduced in Definition 2.15 below, Theorem 2.13 immediately gives inequalities for several surrogates of interesting learning scenarios. For some examples we refer to Theorem 4.3, Example 4.5, Remark 4.8, Theorem 4.20, and Theorem 4.24.

Finally, the almost trivial case of Theorem 2.13, i.e. the case of $\alpha \in \mathcal{A}$ satisfying $\mathcal{R}_{L,P}(\alpha) < \infty$, is (up to measurability) also a direct consequence of the more general [36, Theorem 24] which in turn was inspired by a less general result of [1]. We added the short proof for this case for the sake of completeness.

If the function δ in the above theorem satisfies $\delta(\varepsilon) > 0$ for all $\varepsilon > 0$ then condition (12) is a *uniform* version of the notion of calibration. Let us describe this situation in the following definition:

Definition 2.15 (Uniform calibration) *Let $\mathcal{Q}$ be a set of distributions on Y, and $L_1 : X \times Y \times \mathcal{A} \to [0,\infty)$, $L_2 : X \times Y \times \mathcal{A} \to [0,\infty)$ be two cost functions. We say that* L_2 is uniformly L_1-calibrated with respect to $\mathcal{Q}$ *if for all $\varepsilon > 0$ there exists a $\delta > 0$ such that $\mathcal{M}_{L_2,Q,x}(\delta) \subset \mathcal{M}_{L_1,Q,x}(\varepsilon)$ for all $Q \in \mathcal{Q}$ and $x \in X$, i.e. if for all $Q \in \mathcal{Q}$, $x \in X$, and $\alpha \in \mathcal{A}$ we have*

$$\mathcal{C}_{L_2,Q,x}(\alpha) < \mathcal{C}^*_{L_2,Q,x} + \delta \qquad \Longrightarrow \qquad \mathcal{C}_{L_1,Q,x}(\alpha) < \mathcal{C}^*_{L_1,Q,x} + \varepsilon\,.$$

Furthermore, we say that L_2 is uniformly L_1-calibrated with respect to a distribution P on $X \times Y$ if L_2 is uniformly L_1-calibrated with respect to the set $\{P(.|x) : x \in X\}$.

Obviously the above definition guarantees that L_1 is P-minimizable whenever L_2 is so. Now observe that the larger the function δ satisfying (12) is, the better inequality (13) becomes. Given a set $\mathcal{Q}$ of distributions on Y let us consequently consider the best possible candidate function, that is

$$\delta_{\max}(\varepsilon, \mathcal{Q}) := \sup\Big\{\delta \geq 0 : \mathcal{M}_{L_2,Q,x}(\delta) \subset \mathcal{M}_{L_1,Q,x}(\varepsilon) \text{ for all } x \in X,\, Q \in \mathcal{Q}\Big\}\,, \qquad \varepsilon \in [0,\infty]. \quad (14)$$

Obviously, L_2 is uniformly L_1-calibrated with respect to $\mathcal{Q}$ if and only if $\delta_{\max}(\varepsilon, \mathcal{Q}) > 0$ for all $\varepsilon > 0$.

For a given distribution P on $X \times Y$ let us write $\delta_{\max}(\varepsilon, P) := \delta_{\max}(\varepsilon, \{P(.|x) : x \in X\})$. Assuming that P is of type $\mathcal{Q}$ we then have $\delta_{\max}(\varepsilon, \mathcal{Q}) \leq \delta_{\max}(\varepsilon, P)$ and consequently, the next lemma shows that the distribution-independent function $\delta_{\max}(., \mathcal{Q})$ can be used in Theorem 2.13.[2] Furthermore, it provides a simple method to calculate $\delta_{\max}(., \mathcal{Q})$:

Lemma 2.16 *Let $\mathcal{Q}$ be a set of distributions on Y, and $L_1 : X \times Y \times \mathcal{A} \to [0,\infty)$, $L_2 : X \times Y \times \mathcal{A} \to [0,\infty)$ be two cost functions. Then for all $x \in X$ and all $\varepsilon > 0$ we have*

$$\mathcal{M}_{L_2,Q,x}(\delta_{\max}(\varepsilon, \mathcal{Q})) \subset \mathcal{M}_{L_1,Q,x}(\varepsilon)$$

and

$$\delta_{\max}(\varepsilon, \mathcal{Q}) = \inf_{\substack{Q \in \mathcal{Q} \\ x \in X}} \delta_{\max}\left(\varepsilon, Q, x\right)\,.$$

[2] In Theorem 3.6 we will see that under rather natural conditions, using $\delta_{\max}(., \mathcal{Q})$ does not give worse inequalities than using $\delta_{\max}(\varepsilon, P)$.

Theorem 2.13 states that for distributions P of type $\mathcal{Q}$ we can find inequalities between the excess risks of uniformly calibrated cost functions. The following theorem shows that the uniform calibration is also *necessary* for such inequalities.

Theorem 2.17 *Let $L_1 : X \times Y \times \mathcal{A} \to [0,\infty)$ and $L_2 : X \times Y \times \mathcal{A} \to [0,\infty)$ be two cost functions, and $\mathcal{Q}$ be class of distributions on Y with $\mathcal{C}^*_{L_1,Q,x} < \infty$ and $\mathcal{C}^*_{L_2,Q,x} < \infty$ for all $x \in X$, $Q \in \mathcal{Q}$. Furthermore, let $\delta : [0,\infty] \to [0,\infty]$ be an increasing function with $\delta(0) = 0$ and $\delta(\varepsilon) > 0$ for all $\varepsilon > 0$. If for all distributions P of type $\mathcal{Q}$ satisfying $\mathcal{R}^*_{L_1,P} < \infty$ and $\mathcal{R}^*_{L_2,P} < \infty$, and all measurable $\alpha \in \mathcal{A}$ we have*

$$\delta\big(\mathcal{R}_{L_1,P}(\alpha) - \mathcal{R}^*_{L_1,P}\big) \;\le\; \mathcal{R}_{L_2,P}(\alpha) - \mathcal{R}^*_{L_2,P}\,,$$

then L_2 is uniformly L_1-calibrated with respect to $\mathcal{Q}$.

Unfortunately, it will turn out that in many situations such as regression or density estimation we cannot find interesting, uniformly calibrated surrogates. In such situations the following theorem which presents inequalities between excess risks under less restrictive calibration assumptions can help:

Theorem 2.18 *Let P be a distribution on $X \times Y$ and $L_1 : X \times Y \times \mathcal{A} \to [0,\infty)$, $L_2 : X \times Y \times \mathcal{A} \to [0,\infty)$ be P-minimizable cost functions. Assume that there exist $p \in (0,\infty]$, $q \ge \frac{p+1}{p}$, and a function $b : X \to [0,\infty]$ such that*

$$\delta_{\max}(\varepsilon, P(.|x), x) \ge \varepsilon^q b(x)\,, \qquad\qquad \varepsilon > 0,\, x \in X \tag{15}$$

and $b^{-1} \in \mathcal{L}_p(P_X)$. Then for all $\alpha \in \mathcal{A}$ we have

$$\mathcal{R}_{L_1,P}(\alpha) - \mathcal{R}^*_{L_1,P} \le \|b^{-1}\|^{\frac{1}{q}}_{\mathcal{L}_p(P_X)} \Big(\mathcal{R}_{L_2,P}(\alpha) - \mathcal{R}^*_{L_2,P}\Big)^{\frac{1}{q}}\,.$$

Remark 2.19 The above condition $b^{-1} \in \mathcal{L}_p(P_X)$ measures how much the calibration function $\delta_{\max}(\varepsilon, P(.|x), x)$ violates a uniform lower bound of the form $\delta_{\max}(\varepsilon, P(.|x), x) \ge c\varepsilon^q$, $\varepsilon \in [0,\infty]$. Indeed, the larger we can choose p in Condition (15) the more the shape of b is away from the critical level 0, and thus the closer Condition (15) is to a uniform lower bound. In the extremal case $p = \infty$, Condition (15) actually becomes a uniform bound, and the inequality of Theorem 2.18 equals the inequality of Theorem 2.13.

Finally, it is interesting to note that up to measurability considerations Theorem 2.18 can also be derived from the more general [36, Theorem 24]. Since the direct proof is very easy we decided to include it for the sake of completeness.

3 Loss Functions

In the previous section we have not specified the hypothesis set $\mathcal{A}$ and how the cost function acts on this set. In this subsection we will now investigate *loss functions* which are cost functions where $\mathcal{F}$ consists of measurable functions $f : X \to \mathcal{T}$ and the cost acts on this set pointwise, i.e. in the form $L(x, y, f(x))$. Obviously, all results of the previous section can be applied to such cost functions, but it will turn out in this section that in some cases we can actually show more. The rest of this section is organized as follows: In the first subsection we will formally introduce loss functions. Then our main focus in this subsection is to show that under mild assumptions the conditions of the last section ensuring measurability of e.g. the inner Bayes risks are automatically satisfied. In addition, we show that for loss functions Theorem 2.7 is in some sense optimal. In the following two subsections we then prove some additional properties of certain types of loss functions. Finally, we investigate in the last subsection whether and in which sense approximate L-risk minimizers approximate exact minimizer of the L-risk.

3.1 Measurability Considerations for Loss Functions

In this subsection we first introduce loss functions and some notations related to them. Our main results of this subsection are then presented in Theorem 3.2 and 3.3 which mainly deal with measurability questions.

Let us begin with the following fundamental definition which introduces loss functions:

Definition 3.1 (Loss function) *Let $\mathcal{T}$ be a Polish space. Then a function $L : X \times Y \times \mathcal{T} \to [0,\infty]$ is called a* loss function *if it is measurable.*

Obviously, every loss function L is a cost function, but in general the converse is not true because of the stronger measurability condition loss functions have to satisfy. Moreover, a loss function L also induces a cost function acting on the space $\mathcal{M}(X,\mathcal{T})$ of all measurable functions $f : X \to \mathcal{T}$. Indeed, for $\mathcal{A} := \mathcal{M}(X,\mathcal{T})$ the mapping

$$\begin{aligned} \hat{L} : X \times Y \times \mathcal{A} &\to [0,\infty) \\ (x,y,f) &\mapsto L(x,y,f(x)) \end{aligned}$$

is a cost function since $(x,y) \mapsto L(x,y,f(x))$ is measurable for all $f \in \mathcal{A}$. Now note that the inner risks of L and $\hat{L}$ are related by

$$\mathcal{C}_{\hat{L},Q,x}(f) = \int_Y L(x,y,f(x))\, dQ(x) = \mathcal{C}_{L,Q,x}(f(x))\,, \tag{16}$$

and consequently we have $\mathcal{C}^*_{\hat{L},Q,x} = \mathcal{C}^*_{L,Q,x}$. In the following we are *only* interested in the excess risks of the induced cost function $\hat{L}$, and therefore we write in a slight abuse of notations

$$\mathcal{R}_{L,P}(f) := \mathcal{R}_{\hat{L},P}(f)\,, \qquad\qquad f \in \mathcal{M}(X,\mathcal{T})\,,$$

and analogously, we define $\mathcal{R}^*_{L,P} := \mathcal{R}^*_{\hat{L},P}$. Now recall that all the major results of the previous section required that the involved cost functions are P-minimizable. Having a loss function L we consequently need to ensure that its induced cost function $\hat{L}$ is P-minimizable. This is done in the following lemma whose main difficulty is to ensure the measurability statements.

Theorem 3.2 *Let $L : X \times Y \times \mathcal{T} \to [0,\infty]$ be a loss function and P be a distribution on $X \times Y$. Then the following statements are true:*

i) $\hat{L}$ is P-minimizable if and only if $\mathcal{C}^*_{L,P(.|x),x} < \infty$ and all $x \in \mathcal{X}$.

ii) If we have $\mathcal{M}_{L,P(.|x),x}(0^+) \neq \emptyset$ for all $x \in \mathcal{X}$ then there exists a measurable function $f^*_{L,P} : X \to \mathcal{T}$ with $\mathcal{C}_{L,P(.|x),x}(f^*_{L,P}(x)) = \mathcal{C}^*_{L,P(.|x),x}$ for all $x \in X$, and consequently we have

$$\mathcal{R}_{L,P}(f^*_{L,P}) = \mathcal{R}^*_{L,P}\,.$$

In combination with Lemma 2.5 the above theorem shows that the inner Bayes risks are measurable under rather general conditions, namely the completeness of $(X,\mathcal{X})$. Consequently, one can typically avoid measurability considerations when dealing with loss functions, and therefore earlier investigations on surrogate loss functions (which all avoided measurability considerations) are justified with hindsight.

Now recall the key quantity of the previous section was the calibration function. Since we are only interested in the excess risks of the cost functions $\hat{L}_1$ and $\hat{L}_2$ induced by the loss functions L_1

and L_2 we need to calculate the calibration function of $(\hat{L}_1, \hat{L}_2)$. To this end assume first that we have $\mathcal{C}^*_{L_2,Q,x} < \infty$. Then Equation (16) gives

$$\delta_{\max,L_1,L_2}(\varepsilon, Q, x) = \inf_{\substack{t\in\mathcal{T} \\ t\notin\mathcal{M}_{L_1,Q,x}(\varepsilon)}} \mathcal{C}_{L_2,Q,x}(t) - \mathcal{C}^*_{L_2,Q,x} = \delta_{\max,\hat{L}_1,\hat{L}_2}(\varepsilon, Q, x)\,,$$

where in the last equality we used that $f :\equiv t$ is a measurable function with $f(x) = t$. Moreover, if $\mathcal{C}^*_{L_2,Q,x} = \infty$ the above observation is also true, and consequently it suffices to investigate the calibration function $\delta_{\max,L_1,L_2}(\varepsilon, Q, x)$ when dealing with the pair $(\hat{L}_1, \hat{L}_2)$.

We have already seen in part *i)* of Lemma 2.9 that the calibration function satisfies Condition (6), however, in order to use the calibration function in Theorem 2.8 we also need to know that it is measurable. The following theorem completely resolves this issue, and in addition it *characterizes* the distributions for which implication (7) hold:

Theorem 3.3 *Let $L_1 : X \times Y \times \mathcal{T} \to [0,\infty]$ and $L_2 : X \times Y \times \mathcal{T} \to [0,\infty]$ be loss functions, and P be a distribution on $X \times Y$ such that $\mathcal{R}^*_{L_1,P} < \infty$ and $\mathcal{R}^*_{L_2,P} < \infty$. Then*

$$x \mapsto \delta_{\max}(\varepsilon, P(.|x), x)$$

is measurable for all $\varepsilon > 0$. In addition, consider the following statements:

i) For all $\varepsilon \in (0,\infty]$ we have $P_X\big(\{x \in X : \delta_{\max}(\varepsilon, P(.|x), x) = 0\}\big) = 0$.

ii) For all $\varepsilon \in (0,\infty]$ there is a $\delta > 0$ such that for all measurable functions $f : X \to \mathcal{T}$ we have

$$\mathcal{R}_{L_2,P}(f) < \mathcal{R}^*_{L_2,P} + \delta \qquad \Longrightarrow \qquad \mathcal{R}_{L_1,P}(f) < \mathcal{R}^*_{L_1,P} + \varepsilon\,. \tag{17}$$

Then we have ii) $\Rightarrow$ i), and the inverse implication i) $\Rightarrow$ ii) holds if there exists a function $b \in \mathcal{L}_1(P_X)$ satisfying (6).

The above theorem shows that the L_1-calibration is *necessary* for L_2 being a reasonable surrogate loss in the sense of Question 1. Consequently, the calibration function will be our major tool when investigating specific learning problems in Section 4.

3.2 Supervised Loss Functions

General loss functions can explicitly depend on the input variable x, and hence so do the derived quantities like the inner risks. However, many important loss functions are actually independent of x, and since the theory becomes substantially simpler for such losses we now briefly consider them. Let us begin with the following definition:

Definition 3.4 (Supervised loss functions) *Let $\mathcal{T}$ be a Polish space. A function $L : Y \times \mathcal{T} \to [0,\infty]$ is called a supervised loss function if it is measurable.*

Formally, supervised loss functions are not loss functions, however, it is obvious that every supervised loss function L induces a loss function $\bar{L}$ via $\bar{L}(x, y, t) := L(y, t)$. In the following we *always* identify L with $\bar{L}$. Now note that the inner risk of $\bar{L}$ is

$$\mathcal{C}_{\bar{L},Q,x}(t) = \int_Y L(y,t)dQ(y)\,,$$

i.e. it is *independent of* x. Moreover, the same is obviously true for the derived quantities $\mathcal{C}^*_{L,Q,x}$, $\mathcal{M}_{L,Q,x}(\varepsilon)$, and therefore we usually use the shorter notations

$$\mathcal{C}_{L,Q}(t) := \mathcal{C}_{\bar{L},Q,x}(t), \qquad \mathcal{C}^*_{L,Q} := \mathcal{C}^*_{\bar{L},Q,x}, \quad \text{and} \quad \mathcal{M}_{L,Q}(\varepsilon) := \mathcal{M}_{L,Q,x}(\varepsilon),$$

where $x \in X$ is an arbitrary element. Furthermore, note that if we have two supervised loss functions then the corresponding calibration function $\delta_{\max}(\varepsilon, Q, x)$ is independent of x, too, and hence we analogously write $\delta_{\max}(\varepsilon, Q) := \delta_{\max}(\varepsilon, Q, x)$ for some $x \in X$. Finally note that the often imposed condition $\mathcal{C}^*_{\bar{L},Q,x} < \infty$ is also independent of x. This justifies the following definition:

Definition 3.5 *Let $\mathcal{Q}$ be a class of distributions on Y and $L : Y \times \mathcal{T} \to [0, \infty)$ be a supervised loss function. Then we write*

$$\mathcal{Q}(L) := \left\{ Q \in \mathcal{Q} : \mathcal{C}^*_{L,Q} < \infty \right\}.$$

Obviously, all theorems we have formulated for cost or loss functions also hold for supervised loss functions. In particular, for uniformly calibrated supervised loss functions Theorem 2.13 can be used to establish inequalities between the corresponding excess risks. Now note that for distributions P of type $\mathcal{Q}$ we always have $\delta_{\max}(\varepsilon, P) \leq \delta_{\max}(\varepsilon, \mathcal{Q})$, and consequently one may think that using $\delta_{\max}(\varepsilon, P)$ instead of $\delta_{\max}(\varepsilon, \mathcal{Q})$ leads to sharper inequalities in Theorem 2.13. The following result shows that this intuition is usually false if the set $\mathcal{Q}$ is not chosen overly conservative:

Theorem 3.6 *Let $L_1 : Y \times \mathcal{T} \to [0, \infty]$, $L_2 : Y \times \mathcal{T} \to [0, \infty]$ be supervised loss functions and $\mathcal{Q}$ be a set of distributions on Y with $\mathcal{Q} = \mathcal{Q}(L_2) = \mathcal{Q}(L_1)$. Furthermore assume that there is a distribution μ on X for which there exist mutually disjoint subsets $A_n \subset X$, $n \in \mathbb{N}$, with $\mu(A_n) > 0$ for all $n \in \mathbb{N}$. Then there exists a distribution P of type $\mathcal{Q}$ with $P_X = \mu$ such that for all $\varepsilon > 0$ we have*

$$\delta_{\max}(\varepsilon, P) = \delta_{\max}(\varepsilon, \mathcal{Q}). \tag{18}$$

3.3 Unsupervised Loss Functions

In the previous part of the work we considered loss functions whose inner risks are independent of x. Formally we can also consider loss functions whose inner risks are independent of Q. Before we motivate such losses let us first make a precise definition:

Definition 3.7 (Unsupervised loss functions) *Let $\mathcal{T}$ be a Polish space. A function $L : X \times \mathcal{T} \to [0, \infty]$ is called an unsupervised loss function if it is measurable.*

Like supervised loss functions, unsupervised loss functions are not loss functions. However, it is obvious that every unsupervised loss function L induces a loss function $\bar{L}$ via $\bar{L}(x, y, t) := L(x, t)$. In the following we *always* identify L with $\bar{L}$. Now note that the inner risk of $\bar{L}$ is

$$\mathcal{C}_{\bar{L},Q,x}(t) = L(x, t),$$

i.e. it is *independent of* Q. Moreover, the quantities $\mathcal{C}^*_{L,Q,x}$, $\mathcal{M}_{L,Q,x}(\varepsilon)$ obviously share this property. In the following we thus use the shorthands

$$\mathcal{C}_{L,x}(t) := \mathcal{C}_{L,Q,x}(t), \qquad \mathcal{C}^*_{L,x} := \mathcal{C}^*_{L,Q,x}, \quad \text{and} \quad \mathcal{M}_{L,x}(\varepsilon) := \mathcal{M}_{L,Q,x}(\varepsilon),$$

where Q is an arbitrary distribution. Note that these definitions in particular give

$$\mathcal{C}^*_{L,x} = \inf_{t \in \mathcal{T}} L(x, t). \tag{19}$$

At a first glance it seems rather odd to consider unsupervised loss functions since their associated (inner) risks $\mathcal{C}_{L,P(.|x),x}(.)$ do not depend on the conditional probabilities $P(.|x)$, and hence they apparently do not define a reasonable learning goal. However, we will see later that the key idea of unsupervised loss functions is to encode the dependence of $P(.|x)$ on the first input variable x of L, so that we actually have the possibility to encode dependencies that are *not* of the form of an inner risk. This fact will be particularly important when considering unsupervised learning goals such as density level detection and density estimation. In addition unsupervised loss functions will also serve as a powerful technical tool when considering general questions on loss functions in e.g. Subsection 3.4. In this regard the following class of unsupervised loss functions will turn out to be of particular interest:

Definition 3.8 (Detection loss functions) *Let $A \subset X \times \mathcal{T}$ be a measurable subset and $h : X \to [0,\infty)$ be a measurable function. Then we call $L : X \times \mathcal{T} \to [0,\infty)$ a* detection loss function with respect to (A,h) *if*

$$L(x,t) = \mathbf{1}_A(x,t)\, h(x)\,, \qquad\qquad x \in X\,, t \in \mathcal{T}.$$

Every detection loss function is obviously measurable and hence an unsupervised loss function. In addition, for $x \in X$ and $t \in \mathcal{T}$ our above notations immediately show

$$\mathcal{C}_{L,x}(t) - \mathcal{C}^*_{L,x} = \begin{cases} 0 & \text{if } A(x) := \{t' \in \mathcal{T} : (x,t') \in A\} = \mathcal{T} \\ \mathbf{1}_A(x,t)h(x) & \text{else.} \end{cases} \tag{20}$$

If the detection loss L_2 is uniformly L_1-calibrated with respect to P, then it is straightforward to establish inequalities by Theorem 2.13. However, using the specific form of detection losses one can often improve these inequalities as we will discuss after the following rather general theorem:

Theorem 3.9 *Let $L_1 : X \times \mathcal{T} \to [0,\infty)$ be a detection loss function with respect to (A,h), $L_2 : X \times Y \times \mathcal{T} \to [0,\infty]$ be a loss function, and P be a distribution on $X \times Y$ with $\mathcal{R}^*_{L_1,P} < \infty$, $\mathcal{R}^*_{L_2,P} < \infty$. For $s > 0$ we write*

$$B(s) := \Big\{ x \in X : A(x) \neq \mathcal{T} \text{ and } \delta_{\max}\big(h(x), P(.|x), x\big) < s\, h(x) \Big\}.$$

If there are constants $c > 0$ and $\alpha \in (0,\infty]$ with

$$\int \mathbf{1}_{B(s)} h \, dP_X \;\leq\; (c\,s)^\alpha \tag{21}$$

for all $s > 0$, then for all measurable $f : X \to \mathcal{T}$ we have

$$\mathcal{R}_{L_1,P}(f) - \mathcal{R}^*_{L_1,P} \;\leq\; 2\, c^{\frac{\alpha}{\alpha+1}} \left(\mathcal{R}_{L_2,P}(f) - \mathcal{R}^*_{L_2,P}\right)^{\frac{\alpha}{\alpha+1}}.$$

The above theorem can be used in various settings, but for brevity's sake we only refer to Remark 3.20, Remark 4.4, and Theorem 4.33. Moreover, we have already indicated that Theorem 3.9 improves the inequalities we obtained for general target losses L_1 in various cases. The following two remarks illustrate this:

Remark 3.10 Let L_1 be a detection loss with $h = \mathbf{1}_X$, and assume that Condition (15) is satisfied for some $b : X \to [0,\infty]$ with $b^{-1} \in \mathcal{L}_p(P_X)$ and $q \geq \frac{p+1}{p}$. Then Theorem 2.18 gives

$$\mathcal{R}_{L_1,P}(f) - \mathcal{R}^*_{L_1,P} \;\leq\; \|b^{-1}\|^{\frac{1}{q}}_{\mathcal{L}_p(P_X)} \left(\mathcal{R}_{L_2,P}(f) - \mathcal{R}^*_{L_2,P}\right)^{\frac{1}{q}}. \tag{22}$$

However, we also have $B(s) \subset \{x \in X : b(x) < s\}$, and since $b^{-1} \in \mathcal{L}_p(P_X)$ implies $P_X(\{x \in X : b(x) < s\}) \leq \big(\|b^{-1}\|_p\, s\big)^p$, we find (21) for $c := \|b^{-1}\|_{\mathcal{L}_p(P_X)}$ and $\alpha := p$. Consequently, Theorem 3.9 yields

$$\mathcal{R}_{L_1,P}(f) - \mathcal{R}^*_{L_1,P} \;\leq\; 2\,\|b^{-1}\|^{\frac{p}{p+1}}_{\mathcal{L}_p(P_X)} \big(\mathcal{R}_{L_2,P}(f) - \mathcal{R}^*_{L_2,P}\big)^{\frac{p}{p+1}}\,. \tag{23}$$

Now recall that Theorem 2.18 required $\frac{1}{q} \leq \frac{p}{p+1}$, and therefore (23) is sharper than (22) whenever the excess risk $\mathcal{R}_{L_2,P}(f) - \mathcal{R}^*_{L_2,P}$ is sufficiently small and $q > 1 + 1/p$.

Remark 3.11 Let L_1 be a detection loss and L_2 be a loss that is uniformly L_1-calibrated with respect to some class $\mathcal{Q}$ of distributions. If $\delta_{\max}(., \mathcal{Q}) \geq \varepsilon^q$ for some $q > 1$ and all $\varepsilon \geq 0$ then Theorem 2.13 gives

$$\mathcal{R}_{L_1,P}(f) - \mathcal{R}^*_{L_1,P} \;\leq\; \big(\mathcal{R}_{L_2,P}(f) - \mathcal{R}^*_{L_2,P}\big)^{\frac{1}{q}} \tag{24}$$

for all measurable $f : X \to \mathbb{R}$. However, it holds $B(s) \subset \{x \in X : 0 < h(x) < s^{1/(q-1)}\}$, and consequently, if we have

$$P_X\big(\{x \in X : 0 < h(x) < s\}\big) \;\leq\; (C\,s)^\beta \tag{25}$$

for some constants $C > 0$, $\beta \in (0, \infty]$, and all $s > 0$, then it is easy to check that (21) is satisfied for $c = C^{\frac{\beta q - \beta}{\beta+1}}$ and $\alpha := \frac{\beta+1}{q-1}$. Theorem 3.9 then yields

$$\mathcal{R}_{L_1,P}(f) - \mathcal{R}^*_{L_1,P} \;\leq\; 2\,C^{\frac{\beta q-\beta}{\beta+q}} \big(\mathcal{R}_{L_2,P}(f) - \mathcal{R}^*_{L_2,P}\big)^{\frac{\beta+1}{\beta+q}}\,. \tag{26}$$

Now note that we always have $\frac{\beta+1}{\beta+q} > \frac{1}{q}$, and thus (26) is sharper than (24) whenever the excess risk $\mathcal{R}_{L_2,P}(f) - \mathcal{R}^*_{L_2,P}$ is sufficiently small.

3.4 Self-calibrated Loss Functions

Given a loss function L and a distribution P such that an *exact* minimizer $f^*_{L,P}$ of $\mathcal{R}_{L,P}(.)$ exists one may ask whether, and in which sense, ε-approximate minimizers f_ε of $\mathcal{R}_{L,P}(.)$ approximate $f^*_{L,P}$. The goal of this subsection is to provide some general answers to these questions.

Interestingly, this question and its answers are important for both practical applications of learning algorithms as well as for theoretical considerations. For example, in binary classification (see the next section for a formal definition of this learning problem) one is often not only interested in finding a good classifier f but also an estimate of the conditional probability $P(y = 1|x)$. Now assume that one has obtained a classifier f by using a learning algorithm which is L-risk consistent for a suitable surrogate L of the classification loss. If the minimizer $f^*_{L,P}$ of this loss has a one-to-one correspondence to the conditional probability and we have a positive answer to the above question one can then use a suitable transformation of $f(x)$ to estimate $P(y = 1|x)$. An important non-trivial example of such a loss function is discussed in Example 4.5. Moreover, by combining the results of this subsection with (41) and Example 4.22 one can for example address the question in which sense the absolute distance loss can be used to estimate the (conditional) median in a regression problem. Finally, some positive insight into the above question can also give powerful tools to establish variance bounds which recently turned out to be important for establishing "fast rates" for learning algorithms (see Remark 4.26 for some further discussion).

In the hope that we have convinced the reader of the importance of the above question we now begin with the following definition which introduces another important class of loss functions:

Definition 3.12 (Template loss) *Let $\mathcal{Q}$ be a class of distributions on Y. Then we call a function $L : \mathcal{Q} \times \mathbb{R} \to [0, \infty]$ a* template loss function *if for all $\mathcal{Q}$-type distributions P on $X \times Y$ the P-*instance L_P of L *defined by*

$$\begin{aligned} L_P : X \times \mathbb{R} &\to [0, \infty] \\ (x, t) &\mapsto L\big(P(.|x), t\big)\,, \end{aligned} \tag{27}$$

is measurable.

Again, the main condition of the above definition is the *measurability* which enables us to interpret the P-instances of L as unsupervised loss functions. In particular, we can define the risk of a template loss function by the risk of its P-instance, namely

$$\mathcal{R}_{L,P}(f) := \mathcal{R}_{L_P,P}(f) = \int_X L\big(P(.|x), f(x)\big)\, dP_X(x)\,,$$

where $f : X \to \mathbb{R}$ is a measurable function. Consequently, we define the inner risks of a template loss $L : \mathcal{Q} \times \mathbb{R} \to [0,\infty]$ analogously to the inner risks of unsupervised losses, i.e. we write

$$\mathcal{C}_{L,Q}(t) := L(Q,t), \qquad \text{and} \qquad \mathcal{C}^*_{L,Q} := \inf_{t' \in \mathbb{R}} L(Q,t')$$

for $Q \in \mathcal{Q}$, $t \in \mathbb{R}$. Note that the right hand sides of these definitions have the form we used for unsupervised losses in the sense that no integrals occur while the left hand sides have the form we obtained for supervised losses in the sense that the inner risks are independent of x.

Having defined the inner risks we write, as usual, $\mathcal{M}_{L,Q}(\varepsilon) := \{t \in \mathbb{R} : \mathcal{C}_{L,Q}(t) < \mathcal{C}^*_{L,Q} + \varepsilon\}$, $Q \in \mathcal{Q}$, $\varepsilon > 0$, for the corresponding sets of approximative minimizers. Moreover, given a supervised surrogate loss L_2 we define the calibration function $\delta_{\max}(.,Q) : [0,\infty] \to [0,\infty]$ by

$$\delta_{\max}(\varepsilon, Q) := \delta_{\max,L,L_2}(\varepsilon, Q) := \inf_{\substack{t \in \mathcal{T} \\ t \notin \mathcal{M}_{L,Q}(\varepsilon)}} \mathcal{C}_{L_2,Q}(t) - \mathcal{C}^*_{L_2,Q}\,, \qquad \varepsilon \in [0,\infty],$$

if $\mathcal{C}^*_{L_2,Q} < \infty$, and by $\delta_{\max}(\varepsilon, Q) := \infty$ otherwise. Since in the proofs of Lemma 2.9 and Lemma 2.16 we have not used that the inner risks are defined by integrals, it is then not hard to see that both lemmas also hold for the above calibration function. Consequently, we say that L_2 is L-calibrated with respect to $\mathcal{Q}$ if

$$\delta_{\max}(\varepsilon, Q) > 0$$

for all $\varepsilon > 0$ and $Q \in \mathcal{Q}$. Analogously, we say that L_2 is uniformly L-calibrated with respect to $\mathcal{Q}$ if $\delta_{\max}(\varepsilon, \mathcal{Q}) := \inf_{Q \in \mathcal{Q}} \delta_{\max}(\varepsilon, Q) > 0$ for all $\varepsilon > 0$. If we now consider a P-instance L_P of L we immediately obtain

$$\delta_{\max,L_P,L_2}(\varepsilon, P(.|x), x) = \delta_{\max,L,L_2}(\varepsilon, P(.|x))\,, \qquad \varepsilon \in [0,\infty],\, x \in X. \tag{28}$$

In other words, the L-calibration of L_2 can be investigated analogously to *supervised* losses, i.e. in terms of $\mathcal{Q}$, while the resulting calibration statements can be used to determine the relation between the excess L_2-risk and the excess risk of the *unsupervised* loss L_P.

In order to give a first interesting example of template losses let us now turn to our initial question, namely the approximation properties of approximate risk minimizers. To this end let $L : Y \times \mathbb{R} \to [0,\infty]$ be a supervised loss function. We write

$$\mathcal{Q}_{\min}(L) := \big\{Q \,:\, Q \text{ is a distribution on } Y \text{ with } \mathcal{M}_{L,Q}(0^+) \neq \emptyset\big\}\,, \tag{29}$$

i.e. $\mathcal{Q}_{\min}(L)$ contains the distributions on Y whose inner L-risks have at least one *exact* minimizer. Furthermore, for $Q \in \mathcal{Q}_{\min}(L)$ and $t \in \mathbb{R}$ we define

$$\breve{L}(Q,t) \;:=\; \operatorname{dist}\big(t, \mathcal{M}_{L,Q}(0^+)\big) := \inf_{t' \in \mathcal{M}_{L,Q}(0^+)} |t - t'|\,, \tag{30}$$

i.e. $\breve{L}(Q,t)$ measures the distance of t to the set of elements $t' \in \mathbb{R}$ minimizing $\mathcal{C}_{L,Q}(.)$. Now the following fundamental lemma ensures that the above quantity defines a template loss function:

Lemma 3.13 *Let $L : Y \times \mathbb{R} \to [0,\infty]$ be a supervised loss function. Then $\breve{L} : \mathcal{Q}_{\min}(L) \times \mathbb{R} \to [0,\infty)$ defined by (30) is a template loss function.*

It is almost needless to say that the main statement of the above lemma is again the measurability of the instances of $\breve{L}$.

Now note that the definition of $\breve{L}$ immediately gives $\mathcal{C}^*_{\breve{L},Q} = 0$, and therefore we have

$$\mathcal{M}_{\breve{L},Q}(\varepsilon) = \big\{t \in \mathbb{R} : \breve{L}(Q,t) < \varepsilon\big\} = \big\{t \in \mathbb{R} : \exists t' \in \mathcal{M}_{L,Q}(0^+) \text{ with } |t - t'| < \varepsilon\big\} \tag{31}$$

for all $Q \in \mathcal{Q}_{\min}(L)$ and $\varepsilon \in [0,\infty]$. Furthermore, we have already mentioned, that the results of Lemma 2.9 remain true for template losses. By Inequality (9) the *self-calibration function* $\delta_{\max,\breve{L},L}(.,Q)$ which can be computed by

$$\delta_{\max,\breve{L},L}(\varepsilon,Q) \;=\; \inf_{\substack{t\in\mathbb{R} \\ \operatorname{dist}(t,\mathcal{M}_{L,Q}(0^+))\geq\varepsilon}} \mathcal{C}_{L,Q}(t) - \mathcal{C}^*_{L,Q}$$

thus satisfies

$$\delta_{\max,\breve{L},L}(\operatorname{dist}(t,\mathcal{M}_{L,Q}(0^+)),Q) \leq \mathcal{C}_{L,Q}(t) - \mathcal{C}^*_{L,Q}$$

for all $Q \in \mathcal{Q}_{\min}(L)$ and all $t \in \mathbb{R}$. Note that if $\mathcal{M}_{L,Q}(0^+)$ contains a single element t^*_Q then the latter inequality becomes

$$\delta_{\max,\breve{L},L}(|t - t^*_Q|,Q) \leq \mathcal{C}_{L,Q}(t) - \mathcal{C}^*_{L,Q}\,, \tag{32}$$

and consequently, the calibration function quantifies how well an approximate $\mathcal{C}_{L,Q}(.)$-minimizer t approximates the exact minimizer t^*_Q. This motivates the following definition:

Definition 3.14 (Self-calibration) *Let $L : Y \times \mathbb{R} \to [0,\infty]$ be a supervised loss function and $\mathcal{Q} \subset \mathcal{Q}_{\min}(L)$. We say that L* is (uniformly) self-calibrated with respect to $\mathcal{Q}$, *if L is (uniformly) $\breve{L}$-calibrated with respect to $\mathcal{Q}$.*

If L is a supervised convex loss function then $\mathcal{M}_{L,Q}(0^+)$ is an interval, and hence (31) together with the proof of Lemma 2.11 immediately gives the following result:

Lemma 3.15 (Self-calibration of convex losses) *Every supervised convex loss function $L : Y \times \mathbb{R} \to [0,\infty)$ is self-calibrated with respect to $\mathcal{Q}_{\min}(L)$.*

Note that we will see some examples in Section 4.3 showing that in general supervised convex losses are *not uniformly* self-calibrated. In general, we consequently cannot expect strong inequalities in the sense of Theorem 2.13 for the self-calibration problem. However, even the somewhat weak self-calibration established in Lemma 3.15 can be used to get interesting results for convex loss functions. This is discussed in Remark 3.18 which is a consequence of the following proposition showing that for self-calibrated loss functions, approximate risk minimizers approximate the Bayes decision functions:

Theorem 3.16 *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss function that is self-calibrated with respect to some $\mathcal{Q} \subset \mathcal{Q}_{\min}(L)$, and P be a distribution of $\mathcal{Q}$-type with $\mathcal{R}^*_{L,P} < \infty$. Then for all $\varepsilon > 0$ and $\rho > 0$ there exists a $\delta > 0$ such that for all measurable $f : X \to \mathbb{R}$ we have*

$$\mathcal{R}_{L,P}(f) < \mathcal{R}^*_{L,P} + \delta \qquad \Longrightarrow \qquad P_X\big(\big\{x \in X : \breve{L}_P\big(x, f(x)\big) \geq \rho\big\}\big) < \varepsilon\,.$$

Remark 3.17 In [26] a similar though technically more complicated version of the above theorem has already been proved for binary classification problems.

Remark 3.18 Let L be a convex supervised loss such that the sets $\mathcal{M}_{L,P(.|x),x}(0^+)$ are singletons. By Theorem 3.2 there then exists a P_X-almost surely unique minimizer $f^*_{L,P}$ of $\mathcal{R}_{L,P}(.)$, and thus we have $\breve{L}_P(x,t) = |t - f^*_{L,P}(x)|$ for P_X-almost all $x \in X$. Moreover, the assumptions of the above theorem are obviously satisfied, and consequently we obtain $f_n \to f^*_{L,P}$ in probability for all sequences (f_n) with $\mathcal{R}_{L,P}(f_n) \to \mathcal{R}^*_{L,P}$. In other words, approximate risk minimizers approximate the unique risk minimizer $f^*_{L,P}$ in probability.

Of course, the approximation in probability discussed in the previous remark is a somewhat weak notion and therefore let us finally describe situations in which we can replace it by a stronger notion of approximation:

Proposition 3.19 *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss function that is self-calibrated with respect to some $\mathcal{Q} \subset \mathcal{Q}_{\min}(L)$, and P be a distribution of $\mathcal{Q}$-type with $\mathcal{R}^*_{L,P} < \infty$. Assume that there exist $p \in (0,\infty]$, $q > 0$, and a function $b : X \to [0,\infty]$ with $b^{-1} \in \mathcal{L}_p(P_X)$ and*

$$\delta_{\max,\breve{L}_P,L}(\varepsilon, P(.|x), x) \geq \varepsilon^q \, b(x)\,, \qquad \varepsilon > 0,\, x \in X.$$

Then for all measurable $f : X \to \mathbb{R}$ we have

$$\left(\int_X \big(\breve{L}_P(x, f(x))\big)^{\frac{pq}{p+1}} \, dP_X(x)\right)^{\frac{p+1}{pq}} \leq \|b^{-1}\|^{\frac{1}{q}}_{\mathcal{L}_p(P_X)} \left(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}\right)^{\frac{1}{q}}.$$

Remark 3.20 If $\mathcal{R}_{L,P}(.)$ has an almost surely unique minimizer $f^*_{L,P}$ then Proposition 3.19 ensures $\|.\|_{\frac{pq}{p+1}}$-convergence of f_n to $f^*_{L,P}$ whenever we have $\mathcal{R}_{L,P}(f_n) \to \mathcal{R}^*_{L,P}$. Interestingly, if we can only ensure $b^{-1} \in \mathcal{L}_{p,\infty}(P_X)$, where $\mathcal{L}_{p,\infty}(P_X)$ is a Lorentz space (see e.g. [3]), then the norm $\|.\|_{\frac{pq}{p+1}}$ in the above proposition can be replaced by the Lorentz-norm $\|.\|_{\frac{pq}{p+1},\infty}$ if we combine Theorem 3.9 with the proof of Theorem 3.16.

4 Examples

In this section we apply the general theory of Section 2 to various loss functions of common learning scenarios such as (cost-sensitive) binary classification, regression, density level detection, and density estimation.

4.1 Standard Binary Classification

In this subsection we investigate surrogate loss functions for the binary classification problem. Since such classification calibrated loss functions have already been intensively studied by Bartlett et al. in [1], our main aim in this subsection is to briefly discuss the relationship between their results and our general framework. In addition, the concepts developed in this subsection will be needed in the following subsections when dealing with cost-sensitive classification and density level detection.

Let us begin with briefly recalling the standard binary classification problem. To this end let $Y := \{-1, 1\}$ and $\mathcal{T} := \mathbb{R}$ throughout the subsection. Our target loss is the supervised loss function $L_{\text{class}} : Y \times \mathbb{R} \to [0,\infty)$ defined by

$$L_{\text{class}}(y,t) := \mathbf{1}_{(-\infty,0]}\big(y \operatorname{sign} t\big)\,, \qquad y \in Y,\, t \in \mathbb{R}, \tag{33}$$

where we use the convention $\operatorname{sign} 0 := 1$. For a given distribution P on $X \times Y$ an easy computation then shows that $\mathcal{R}_{L_{\text{class}},P}(f)$ is the classification risk discussed in the introduction.

In the following, $\mathcal{Q}_Y$ denotes the set of all distributions on Y. We say that a supervised loss function L is *(uniformly) classification calibrated* if it is (uniformly) L_{class}-calibrated with respect to

$\mathcal{Q}_Y$. Furthermore observe that any distribution $Q \in \mathcal{Q}_Y$ can be uniquely described by an $\eta \in [0,1]$ using the identification $\eta = Q(\{1\})$. If L is a supervised loss function we therefore use the notations

$$\mathcal{C}_{L,\eta}(t) := \mathcal{C}_{L,Q}(t), \qquad \mathcal{C}^*_{L,\eta} := \mathcal{C}^*_{L,Q}, \qquad \mathcal{M}_{L,\eta}(\varepsilon) := \mathcal{M}_{L,Q}(\varepsilon), \qquad \text{and} \qquad \delta_{\max}(\varepsilon,\eta) := \delta_{\max}(\varepsilon, Q)$$

for $t \in \mathbb{R}$ and $\varepsilon > 0$. Now our first aim is to compute $\mathcal{M}_{L_{\text{class}},\eta}(\varepsilon)$ and $\delta_{\max}(\varepsilon,\eta)$:

Lemma 4.1 *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss function. Then for all $\eta \in [0,1]$, $\varepsilon > 0$ we have*

$$\delta_{\max}(\varepsilon,\eta) = \begin{cases} \infty & \text{if } \varepsilon > |2\eta - 1|, \\ \inf_{t \in \mathbb{R}:(2\eta-1)\operatorname{sign} t \le 0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} & \text{if } \varepsilon \le |2\eta - 1|. \end{cases}$$

Proof: For $t \in \mathbb{R}$ a well-known and easy calculation shows

$$\mathcal{C}_{L_{\text{class}},\eta}(t) - \mathcal{C}^*_{L_{\text{class}},\eta} = \big|2\eta - 1\big| \cdot \mathbf{1}_{(-\infty,0]}\big((2\eta-1)\operatorname{sign} t\big)\,. \tag{34}$$

Now if $\varepsilon > |2\eta - 1|$ we obviously have $\mathcal{C}_{L_{\text{class}},\eta}(t) - \mathcal{C}^*_{L_{\text{class}},\eta} < \varepsilon$ for all $t \in \mathbb{R}$ and hence we obtain $\mathcal{M}_{L_{\text{class}},\eta}(\varepsilon) = \mathbb{R}$. On the other hand for $\varepsilon \le |2\eta - 1|$ we have $\mathcal{C}_{L_{\text{class}},\eta}(t) - \mathcal{C}^*_{L_{\text{class}},\eta} < \varepsilon$ if and only if $(2\eta - 1)\operatorname{sign} t > 0$, and hence $\mathcal{M}_{L_{\text{class}},\eta}(\varepsilon) = \{t \in \mathbb{R} : (2\eta-1)\operatorname{sign} t > 0\}$. ∎

In the following we will restrict our considerations to the following class of supervised loss functions used in many classification algorithms.

Definition 4.2 *A supervised loss function $L : Y \times \mathbb{R} \to [0,\infty)$ is called* margin-based *if there exists a $\varphi : \mathbb{R} \to [0,\infty)$ with $L(y,t) = \varphi(yt)$ for all $y \in Y$ and all $t \in \mathbb{R}$.*

Note that the classification loss L_{class} is *not* margin-based, but many other loss functions are. For some examples we refer to Table 1. Now one of the main results on margin-based losses established in [1] reads as follows:

Theorem 4.3 *Let L be a margin-based loss function. Then the following are equivalent:*

i) L is classification calibrated.

ii) L is uniformly classification calibrated.

iii) The function $H : [0,1] \to [0,\infty)$ defined by

$$H(\eta) := \inf_{t \in \mathbb{R}:(2\eta-1)t \le 0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta}, \qquad \eta \in [0,1], \tag{35}$$

satisfies $H(\eta) > 0$ for all $\eta \in [0,1]$ with $\eta \neq 1/2$.

Furthermore, for $\delta : [0,1] \to [0,\infty)$ defined by $\delta(\varepsilon) := H(\frac{1+\varepsilon}{2})$, $\varepsilon \in [0,1]$, we have

$$\delta^{**}(\varepsilon) \le \delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y)\,, \qquad \varepsilon \in [0,1], \tag{36}$$

and both quantities are actually equal whenever L is continuous.

The above theorem shows that for binary classification, calibration coincides with uniform calibration, and consequently a positive answer to Question 1 automatically gives a positive answer to Question 2. Moreover, in order to find the corresponding inequality, it suffices to compute the function H (or the calibration function for continuous surrogates). For many interesting loss functions

Name of loss function	$\varphi(t)$ for $t \in \mathbb{R}$	$H(\eta)$ for $\eta \in [0,1]$	$\delta^{**}(\varepsilon)$ for $\varepsilon \in (0,1]$
Hinge loss	$\max\{0, 1-t\}$	$\lvert 2\eta - 1\rvert$	ε
Truncated least squares	$\bigl(\max\{0,1-t\}\bigr)^2$	$(2\eta-1)^2$	ε^2
Least squares	$(1-t)^2$	$(2\eta-1)^2$	ε^2
Exponential loss	$\exp(-t)$	$1-2\sqrt{\eta(1-\eta)}$	$1-\sqrt{1-\varepsilon^2}$
Sigmoid loss	$1-\tanh(t)$	$\lvert 2\eta-1\rvert$	ε
Logistic loss	$\ln(1+e^{-t})$	$\ln 2 + \eta\ln\eta + (1-\eta)\ln(1-\eta)$	$\frac{(1+\varepsilon)\ln(1+\varepsilon)+(1-\varepsilon)\ln(1-\varepsilon)}{2}$

Table 1: Some common margin-based loss functions and the corresponding values for $H(\eta)$ and δ^{**} defined in Theorem 4.3. All results are taken from [1] besides the ones for the least squares loss and the logistic loss.

this has already been done by Bartlett *et al.* in [1]. Their results together with the corresponding values of the function δ^{**} defined in Theorem 4.3 are summarized in Table 1. Finally, recall that if L is a *convex* margin-based loss function then it was shown in [1] that L is classification calibrated if and only if its associated φ is differentiable at 0 with $\varphi'(0) < 0$. In addition, in this case [1, Theorem 4] shows the simple formula

$$\delta^{**}_{\max}(\varepsilon, \mathcal{Q}_Y) = \varphi(0) - \mathcal{C}^*_{L,\frac{1+\varepsilon}{2}}\,, \qquad \varepsilon \in [0,1].$$

Remark 4.4 It is interesting to note that Equation (34) can be used to describe the classification loss by a detection loss function. Indeed, if for a given distribution P on $X \times Y$ we write

$$L_P(x,t) := \bigl|2\eta(x)-1\bigr| \cdot \mathbf{1}_{(-\infty,0]}\bigl((2\eta(x)-1)\operatorname{sign} t\bigr)\,, \qquad x \in X\, t \in \mathbb{R},$$

then $L_P : X \times \mathbb{R} \to [0,\infty)$ is a detection loss with $h(x) = |2\eta(x)-1|$, $x \in X$, and (34) states

$$\mathcal{C}_{L_{\text{class}},\eta(x)}(t) - \mathcal{C}^*_{L_{\text{class}},\eta(x)} = \mathcal{C}_{L_P,x}(t) - \mathcal{C}^*_{L_P,x}$$

for all $x \in X$, $t \in \mathbb{R}$. Furthermore, Condition (25) is then a weaker version of Tsybakov noise assumption in the sense of [29], and the resulting inequality of Theorem 3.9 essentially coincides with that of [1, Thm. 10]. In this regard it is also interesting to note that the function $\delta^{**}(\varepsilon)$ for the logistic loss (see Table 1) behaves like ε^2, so that for this loss the inequality of Theorem 3.9 can be substantially simplified.

Sometimes, practical classification problems do not only require a small classification risk but also an estimate of the conditional probability $\eta(x) = P(y=1|x)$, $x \in X$. If we have a margin-based loss function L for which there is a one-to-one transformation between the sets of minimizers $\mathcal{M}_{L,\eta}(0^+)$ and η then it seems natural to use self-calibration properties of L to investigate whether suitably transformed approximate L-risk minimizers approximate η. This approach is discussed in the following example:

Example 4.5 (Estimating the conditional probabilities with the logistic loss) Let L_{logist} be the logistic loss defined by $\varphi(t) = \ln(1+e^{-t})$, $t \in \mathbb{R}$. Then it is well-known that

$$\mathcal{M}_{L_{\text{logist}},\eta}(0^+) = \Bigl\{\ln\Bigl(\frac{\eta}{1-\eta}\Bigr)\Bigr\}\,, \qquad \eta \in (0,1).$$

In other words, if t_η denotes the element contained in $\mathcal{M}_{L_{\text{logist}},\eta}(0^+)$ then we have $\eta = \frac{1}{1+e^{-t_\eta}}$. Consequently, if t approximately minimizes $\mathcal{C}_{L_{\text{logist}},\eta}(.)$ then it is close to t_η by Lemma 3.15 and hence $\frac{1}{1+e^{-t}}$ can serve as

an estimate of η. However, investigating the quality of this estimate by the self-calibration function of L_{logist} causes some technical problems since L_{logist} is only self-calibrated with respect to the distributions $Q \in \mathcal{Q}_Y$ with $Q(\{1\}) \neq 0, 1$. Consequently, we now assess the quality of the above estimate *directly*. To this end we define $L : \mathcal{Q}_Y \times \mathbb{R} \to [0, \infty)$ by $L(\eta, t) := \left|\eta - \frac{1}{1+e^{-t}}\right|$, $\eta \in [0, 1]$, $t \in \mathbb{R}$. Then L is template loss function which measures the distance between η and its estimate in the sense of the above discussion. Let us compute the calibration function of (L, L_{logist}). To this end we first observe that $\mathcal{C}^*_{L,\eta} = 0$ for all $\eta \in [0, 1]$, and hence we have $\mathcal{M}_{L,\eta}(\varepsilon) = \{t \in \mathbb{R} : L(\eta, t) < \varepsilon\}$. Consequently, for $\mathcal{C}_\eta(t) := \mathcal{C}_{L_{\text{logist}},\eta}(t) - \mathcal{C}^*_{L_{\text{logist}},\eta}$ Lemma 2.11 gives

$$\delta_{\max,L,L_{\text{logist}}}(\varepsilon, \eta) = \min\left\{\mathcal{C}_\eta\left(\ln\Big(\frac{\eta - \varepsilon}{1 - \eta + \varepsilon}\Big)_+\right), \mathcal{C}_\eta\left(-\ln\Big(\frac{1 - \eta - \varepsilon}{\eta + \varepsilon}\Big)_+\right)\right\},$$

where we use the convention $(x)_+ := \max\{0, x\}$, $x \in \mathbb{R}$, and $\mathcal{C}_\eta(\pm\infty) := \infty$. From this we can easily conclude $\delta_{\max,L,L_{\text{logist}}}(\varepsilon, \eta) = \delta_{\max,L,L_{\text{logist}}}(\varepsilon, 1-\eta)$ for all $\varepsilon \geq 0$, $\eta \in [0, 1]$. Moreover, for fixed $\varepsilon \in (0, 1/2)$ and $\varepsilon < \eta < 1 - \varepsilon$ some calculations show

$$\eta \ln \frac{\eta}{\eta - \varepsilon} + (1 - \eta) \ln \frac{1 - \eta}{1 - \eta + \varepsilon} \;\leq\; \eta \ln \frac{\eta}{\eta + \varepsilon} + (1 - \eta) \ln \frac{1 - \eta}{1 - \eta - \varepsilon}$$

if and only if $\eta \geq \frac{1}{2}$, and consequently for $\eta \in [0, 1/2]$ we find

$$\delta_{\max,L,L_{\text{logist}}}(\varepsilon, \eta) = \begin{cases} \eta \ln \frac{\eta}{\eta+\varepsilon} + (1 - \eta) \ln \frac{1-\eta}{1-\eta-\varepsilon} & \text{if } \varepsilon < 1 - \eta \\ \infty & \text{else.} \end{cases}$$

In order to investigate whether L_{logist} is L-calibrated with respect to $\mathcal{Q}_Y$ let us now find a simple lower bound of the above calibration function. To this end let $h(\eta) := \eta \ln \frac{\eta}{\eta+\varepsilon}$. Then its derivative satisfies

$$h'(\eta) = \ln \frac{\eta}{\eta + \varepsilon} + \frac{\varepsilon}{\eta + \varepsilon} = \ln\Big(1 - \frac{\varepsilon}{\eta + \varepsilon}\Big) + \frac{\varepsilon}{\eta + \varepsilon} \;\leq\; -\frac{\varepsilon}{\eta + \varepsilon} + \frac{\varepsilon}{\eta + \varepsilon} = 0\,,$$

and hence we find $\eta \ln \frac{\eta}{\eta+\varepsilon} \geq \frac{1}{2} \ln \frac{1}{1+2\varepsilon}$ for all $\eta \in [0, 1/2]$, $\varepsilon \geq 0$. Analogously, we obtain $(1 - \eta) \ln \frac{1-\eta}{1-\eta-\varepsilon} \geq \ln \frac{1}{1-\varepsilon}$, for $\eta \in [0, 1/2]$, $\varepsilon \in [0, 1 - \eta)$. Both estimates together then yield

$$\delta_{\max,L,L_{\text{logist}}}(\varepsilon, \eta) \geq \frac{1}{2} \ln \frac{1}{1 + 2\varepsilon} + \ln \frac{1}{1 - \varepsilon} \geq \varepsilon^2$$

for all $\eta \in [0, 1/2]$, $\varepsilon \in [0, 1 - \eta)$. From this we easily can conclude that L_{logist} is *uniformly* L-calibrated with respect $\mathcal{Q}_Y$ with $\delta_{\max,L,L_{\text{logist}}}(\varepsilon, \mathcal{Q}_Y) \geq \varepsilon^2$ for all $\varepsilon \geq 0$. If we now consider the squared version L^2 of L then we obviously have $\delta_{\max,L^2,L_{\text{logist}}}(\varepsilon, \mathcal{Q}_Y) = \delta_{\max,L,L_{\text{logist}}}(\sqrt{\varepsilon}, \mathcal{Q}_Y) \geq \varepsilon$ for all $\varepsilon \geq 0$. For measurable $f : X \to \mathbb{R}$ Theorem 2.13 consequently gives

$$\left(\int_X \Big|\eta(x) - \frac{1}{1 + e^{-f(x)}}\Big|^2 \, dP_X(x)\right)^{1/2} \leq \sqrt{\mathcal{R}_{L_{\text{logist}},P}(f) - \mathcal{R}^*_{L_{\text{logist}},P}}\,,$$

i.e. the we can assess the quality of the estimate $\frac{1}{1+e^{-f(x)}}$ in terms of the $\|.\|_2$-norm.

4.2 Cost-sensitive Binary Classification

In this section we investigate surrogate loss functions for a simple class of cost-sensitive binary classification problems (see [10] for a definition of general cost-sensitive binary classification). Unlike in standard binary classification, in the considered cost-sensitive binary classification scenario the two types of errors are assigned different costs. To make this precise, let $\alpha \in (0, 1)$ be a real number, $Y := \{-1, 1\}$. Then our target loss is the supervised loss function $L_{\alpha\text{-class}} : Y \times \mathbb{R} \to [0, \infty)$ defined by

$$L_{\alpha\text{-class}}(y, t) := \begin{cases} 1 - \alpha & \text{if } y = 1 \text{ and } t < 0 \\ \alpha & \text{if } y = -1 \text{ and } t \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{37}$$

Obviously $L_{\alpha\text{-class}}$ is a supervised loss function and we have $2L_{\frac{1}{2}\text{-class}} = L_{\text{class}}$. As in the standard case we begin with computing the calibration function.

Lemma 4.6 *Let $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss function. Then for all $\alpha \in (0,1)$, $\eta \in [0,1]$ and all $\varepsilon > 0$ we have*

$$\delta_{\max}(\varepsilon,\eta) = \begin{cases} \infty & \text{if } \varepsilon > |\eta - \alpha|, \\ \inf_{t\in\mathbb{R}:(\eta-\alpha)\operatorname{sign} t \le 0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} & \text{if } \varepsilon \le |\eta-\alpha|. \end{cases}$$

Proof: A simple calculation shows $\mathcal{C}_{L_{\alpha\text{-class}},\eta}(t) - \mathcal{C}^*_{L_{\alpha\text{-class}},\eta} = \big|\eta - a\big| \cdot \mathbf{1}_{(-\infty,0]}\big((\eta-\alpha)\operatorname{sign} t\big)$, and thus the assertion follows as in the proof of Lemma 4.1. ∎

Using a weighted version of the standard binary classification loss is a somewhat old idea in machine learning (see e.g. [10]). In the following we will focus on the question whether weighted versions of margin-based losses can serve as a surrogate for the cost-sensitive loss defined by (37). To this end let L be a margin-based loss function with corresponding function $\varphi : \mathbb{R} \to [0,\infty)$. For $\alpha \in (0,1)$ we define the *α-weighted version L_α of L* by

$$L_\alpha(y,t) := \begin{cases} (1-\alpha)\varphi(t) & \text{if } y = 1 \\ \alpha\varphi(-t) & \text{if } y = -1, \end{cases} \qquad t \in \mathbb{R}.$$

Moreover, when considering these weighted versions, we frequently use the quantities

$$w_\alpha(\eta) := (1-\alpha)\eta + \alpha(1-\eta) \qquad \text{and} \qquad \vartheta_\alpha(\eta) := \frac{(1-\alpha)\eta}{(1-\alpha)\eta + \alpha(1-\eta)}$$

which are defined for all $\eta \in [0,1]$. Now we can characterize when weighted versions of margin based loss functions are $L_{\alpha\text{-class}}$-calibrated.

Theorem 4.7 *For a margin-based loss L and $\alpha \in (0,1)$ the following statements are equivalent:*

i) L_α is uniformly $L_{\alpha\text{-class}}$-calibrated with respect to $\mathcal{Q}_Y$.

ii) L_α is $L_{\alpha\text{-class}}$-calibrated with respect to $\mathcal{Q}_Y$.

iii) L is classification calibrated.

iv) The function $H_\alpha : [0,1] \to [0,\infty)$ defined by

$$H_\alpha(\eta) := \inf_{t\in\mathbb{R}:(\eta-\alpha)t\le 0} \mathcal{C}_{L_\alpha,\eta}(t) - \mathcal{C}^*_{L_\alpha,\eta}, \qquad \eta \in [0,1],$$

satisfies $H_\alpha(\eta) > 0$ for all $\eta \in [0,1]$ with $\eta \ne \alpha$.

Furthermore, if one of the statements is true and H is defined by (35) then we have

$$H_\alpha(\eta) = w_\alpha(\eta) H\big(\vartheta_\alpha(\eta)\big)\,. \tag{38}$$

Proof: Some easy calculations show $2\vartheta_\alpha(\eta) - 1 = \frac{\eta-\alpha}{(1-\alpha)\eta+\alpha(1-\eta)}$ and $|\eta-\alpha| \le |2\vartheta_\alpha(\eta)-1|$. Now let $\delta_{\max}(\varepsilon,\eta)$ be defined with respect to L_{class} and L, and $\delta_{\max,\alpha}(\varepsilon,\eta)$ be defined with respect to $L_{\alpha\text{-class}}$ and L_α. Then for $\varepsilon \le |\eta-\alpha|$ our preliminary considerations yield

$$\begin{aligned} \delta_{\max,\alpha}(\varepsilon,\eta) = \inf_{\substack{t\in\mathbb{R} \\ (\eta-\alpha)\operatorname{sign} t \le 0}} \mathcal{C}_{L_\alpha,\eta}(t) - \mathcal{C}^*_{L_\alpha,\eta} &= w_\alpha(\eta) \inf_{\substack{t\in\mathbb{R} \\ (2\vartheta_\alpha(\eta)-1)\operatorname{sign} t \le 0}} \mathcal{C}_{L,\vartheta_\alpha(\eta)}(t) - \mathcal{C}^*_{L,\vartheta_\alpha(\eta)} \\ &= w_\alpha(\eta)\delta_{\max}(\varepsilon,\vartheta_\alpha(\eta))\,. \end{aligned}$$

Name of loss function	$H(\eta)$	$H_\alpha(\eta)$	$\delta_\alpha^{**}(\varepsilon)$
Hinge loss	$\lvert 2\eta - 1\rvert$	$\lvert \eta - \alpha\rvert$	ε
Truncated least squares	$(2\eta-1)^2$	$\frac{(\eta-\alpha)^2}{\alpha+\eta-2\alpha\eta}$	$\frac{\varepsilon^2}{2\alpha(1-\alpha)+\varepsilon(1-2\alpha)}$
Least squares	$(2\eta-1)^2$	$\frac{(\eta-\alpha)^2}{\alpha+\eta-2\alpha\eta}$	$\frac{\varepsilon^2}{2\alpha(1-\alpha)+\varepsilon(1-2\alpha)}$
Exponential loss	$1-2\sqrt{\eta(1-\eta)}$	$\alpha+\eta-2\alpha\eta+2A_0\sqrt{\eta(1-\eta)}$	$2A_0(A_0-A_\varepsilon)+\varepsilon(2\alpha-1)$
Sigmoid loss	$\lvert 2\eta - 1\rvert$	$\lvert \eta - \alpha\rvert$	ε

Table 2: The functions H, H_α and δ_α^{**} for some common margin-based losses. For the exponential loss we used the abbreviation $A_\varepsilon := \sqrt{(\alpha-\varepsilon)(1-\alpha+\varepsilon)}$, $\varepsilon \geq 0$. Furthermore we omitted the logistic loss because the corresponding formulas a too long to fit into the table. Note that for the hinge loss and the sigmoid loss the function δ_α^{**} is actually *independent* of α.

From this we immediately obtain the equivalence of *ii)* and *iii)*. Furthermore, *i)* $\Rightarrow$ *ii)* is trivial, and *iii)* $\Rightarrow$ *i)* follows from $\delta_{\max,\alpha}(\varepsilon,\eta) = w_\alpha(\eta)\delta_{\max}(\varepsilon,\vartheta_\alpha(\eta)) \geq \min\{\alpha,1-\alpha\}\delta_{\max}(\varepsilon,\vartheta_\alpha(\eta))$ and Theorem 4.3. Now, if L is classification calibrated, *iii)* of Theorem 4.3 yields

$$\mathcal{C}_{L_\alpha,\eta}(0) = w_\alpha(\eta)\,\mathcal{C}_{L,\vartheta_\alpha(\eta)}(0) > w_\alpha(\eta)\,\mathcal{C}^*_{L,\vartheta_\alpha(\eta)} = \mathcal{C}^*_{L_\alpha,\eta}\,,$$

and hence Lemma 4.6 together with *ii)* implies $H_\alpha(\eta) > 0$ for all $\eta \neq \alpha$. Conversely, if *iv)* holds then we have $\delta_{\max,\alpha}(\varepsilon,\eta) \geq H_\alpha(\eta) > 0$ for $\eta \neq \alpha$ and $0 < \varepsilon \leq |\eta-\alpha|$, and hence L_a is L_{class}-calibrated with respect to $\mathcal{Q}_Y$. Finally the proof of (38) is analogous to the proof of $\delta_{\max,\alpha}(\varepsilon,\eta) = w_\alpha(\eta)\delta_{\max}(\varepsilon,\vartheta_\alpha(\eta))$. ∎

Remark 4.8 The above results can be used to obtain inequalities between the excess risks of $L_{\alpha\text{-class}}$ and α-weighted versions of margin-based, classification calibrated loss functions. In order to find such inequalities let us write

$$\delta_\alpha(\varepsilon) := \inf_{\substack{\eta\in[0,1]\\ |\eta-\alpha|\geq\varepsilon}} H_\alpha(\eta)\,,$$

where $\alpha_{\max} := \max\{\alpha,1-\alpha\}$ and $\varepsilon \in [0,\alpha_{\max}]$. For $\varepsilon \in [0,\alpha_{\max}]$ Lemma 4.6 then yields

$$\inf_{Q\in\mathcal{Q}_Y} \delta_{\max}(\varepsilon,Q) \;=\; \inf_{\substack{\eta\in[0,1]\\ |\eta-\alpha|\geq\varepsilon}} \;\inf_{\substack{t\in\mathbb{R}\\ (\eta-\alpha)\operatorname{sign} t\leq 0}} \mathcal{C}_{L_\alpha,\eta}(t) - \mathcal{C}^*_{L_\alpha,\eta} \;\geq\; \inf_{\substack{\eta\in[0,1]\\ |\eta-\alpha|\geq\varepsilon}} H_\alpha(\eta)\,,$$

and consequently we have

$$\delta_\alpha^{**}(\varepsilon) \leq \delta_{\max}^{**}(\varepsilon,\mathcal{Q}_Y)\,.$$

Moreover, it is obvious that for continuous L we even have equality in the above formula. Furthermore, for some loss functions we already know $H(\eta)$, $\eta\in[0,1]$, and hence the computation of $\delta_\alpha^{**}(\varepsilon)$ is straightforward. The corresponding results are summarized in Table 2.

Up to now we have only considered α-weighted versions of classification calibrated loss functions to get $L_{\alpha\text{-class}}$-calibrated loss functions. We finally show that this is in some sense the only choice:

Proposition 4.9 *Let $\alpha,\beta\in(0,1)$, L be a margin-based classification calibrated loss function, and L_β be its β-weighted version. Then L_β is $L_{\alpha\text{-class}}$-calibrated if and only if $\beta=\alpha$.*

Proof: We have already seen L_α is $L_{\alpha\text{-class}}$-calibrated, and therefore let us now assume $\alpha \neq \beta$. Without loss of generality we only consider the case $\beta > \alpha$. Then for a fixed $\eta \in (\alpha, \beta)$ an easy computation shows $\vartheta_\beta(\eta) < \frac{1}{2}$, and hence for $\varepsilon > 0$ with $\varepsilon \leq |\eta - \alpha|$ we obtain

$$\delta_{\max}(\varepsilon, \eta) = \inf_{(n-\alpha)\operatorname{sign} t \leq 0} \mathcal{C}_{L_\alpha,\eta}(t) - \mathcal{C}^*_{L_\alpha,\eta} = w_\beta(\eta) \inf_{t<0} \mathcal{C}_{L,\vartheta_\beta(\eta)}(t) - \mathcal{C}^*_{L,\vartheta_\beta(\eta)} \,. \tag{39}$$

Furthermore, since L is classification calibrated we have $\inf_{t\geq 0} \mathcal{C}_{L,\vartheta_\beta(\eta)}(t) - \mathcal{C}^*_{L,\vartheta_\beta(\eta)} > 0$, and since $\inf_{t\in\mathbb{R}} \mathcal{C}_{L,\vartheta_\beta(\eta)}(t) - \mathcal{C}^*_{L,\vartheta_\beta(\eta)} = 0$ we find $\inf_{t<0} \mathcal{C}_{L,\vartheta_\beta(\eta)}(t) - \mathcal{C}^*_{L,\vartheta_\beta(\eta)} = 0$. Together with (39) this shows that L_β is not $L_{\alpha\text{-class}}$-calibrated. ∎

Remark 4.10 Note that the above proposition in particular shows that an α-weighted version of a classification calibrated margin-based loss function is classification calibrated (in the sense of the previous subsection) if and only if $\alpha = \frac{1}{2}$. In other words, changing the weights produces a loss function which is not classification calibrated, and hence this often used heuristic for unbalanced datasets may lead to systematical errors.

4.3 Regression

In this subsection we investigate the use of surrogate loss functions for regression problems. Recall, that the general goal in regression is to predict real-valued outputs, i.e. the label space Y is either the entire real line $\mathbb{R}$ or an interval. The most commonly used loss function for regression is the least squares loss L_{lsquares} defined by $L_{\text{lsquares}}(y,t) := (y-t)^2$. However, this loss is known to be sensitive against outliers, and hence one of our goals is to identify L_{lsquares}-calibrated loss functions. Furthermore, we investigate the problem of finding the mean for symmetric noise distributions and discuss self-calibration for a class of loss functions often employed in regression problems. Let us begin by introducing this type of loss functions:

Definition 4.11 *Let $L : \mathbb{R} \times \mathbb{R} \to [0,\infty)$ be a supervised loss function. We say that L is:*

i) distance-based *if there is a $\psi : \mathbb{R} \to [0,\infty)$ with $\psi(0) = 0$ and $L(y,t) = \psi(y-t)$ for all $y, t \in \mathbb{R}$.*

ii) symmetric *if L is distance-based and its associated ψ satisfies $\psi(r) = \psi(-r)$ for all $r \in \mathbb{R}$.*

Obviously, the least squares loss, and more general, the L_p-loss functions, $p > 0$, defined by $|y-t|^p$ are symmetric. Moreover the logistic loss for regression, Huber's loss and the ϵ-insensitive loss are further examples of symmetric loss functions (see also Table 3).

The following definition introduces some important notions for distance-based loss functions:

Definition 4.12 *Let $p > 0$ and $L : \mathbb{R}\times\mathbb{R} \to [0,\infty)$ be a distance-based loss function with associated $\psi : \mathbb{R} \to [0,\infty)$. We say that L is:*

i) (strictly, uniformly) *convex if ψ is (strictly, uniformly) convex.*

ii) of growth type p *if there are $c_1, c_2 > 0$ with $c_1 r^p \leq \psi(r) \leq c_2 r^p$ for all sufficiently large r.*

iii) locally Lipschitz continuous *if for all $a > 0$ the restriction $\psi_{|[-a,a]}$ is Lipschitz continuous.*

For our analysis we also need some notions related to distributions on $\mathbb{R}$. To this end let Q be a distribution on $Y := \mathbb{R}$ with finite first moment, i.e. $|Q|_1 := \mathbb{E}_{y\sim Q}|y| < \infty$. Then the mean of Q is

$$\mathbb{E}Q := \int_Y y \, dQ(y) \,.$$

We call Q *symmetric* around some center $c \in \mathbb{R}$, if $Q(c+A) = Q(c-A)$ for all measurable $A \subset [0,\infty)$. It is not hard to see that the mean $\mathbb{E}Q$ is the only center whenever $|Q|_1 < \infty$. Furthermore we say that Q is symmetric if it is symmetric around some $c \in \mathbb{R}$. Obviously, Q is symmetric around c if and only if its *centered version* $Q^{(c)}$ defined by $Q^{(c)}(A) := Q(c+A)$, $A \subset \mathbb{R}$ measurable, is centered around 0. In the following we denote the set of distributions with first finite moment by $\mathcal{Q}_{\mathbb{R}}$, and the set of distributions with support contained in the *bounded* interval I by $\mathcal{Q}_I$. Moreover, we write

$$\mathcal{Q}_{\text{bounded}} := \bigcup_{M>0} \mathcal{Q}_{[-M,M]}\,,$$

for the set of distributions with bounded support. Clearly we have $\mathcal{Q}_I \subset \mathcal{Q}_{\text{bounded}} \subset \mathcal{Q}_{\mathbb{R}}$ for all bounded intervals I, and if L is a continuous, distance-based loss function we actually have $\mathcal{Q}_{\text{bounded}} \subset \mathcal{Q}_{\mathbb{R}}(L)$. Moreover, we denote the set of all *symmetric* distributions with first finite moment by $\mathcal{Q}_{\text{sym}}$. Finally, the sets $\mathcal{Q}_{I,\text{sym}}$ for $I \subset \mathbb{R}$ being bounded interval, and $\mathcal{Q}_{\text{bounded,sym}}$ are defined in the obvious way.

Our first goal is to identify distance-based, L_{lsquares}-calibrated loss functions L. To this end recall

$$\mathcal{M}_{L_{\text{lsquares}},Q}(0^+) = \{\mathbb{E}Q\}$$

for all $Q \in \mathcal{Q}(L_{\text{lsquares}})$, and consequently, if L is a L_{lsquares}-calibrated loss function then we must have $\mathcal{M}_{L,Q}(0^+) \subset \{\mathbb{E}Q\}$ for all $Q \in \mathcal{Q}(L)$. This observation motivates the following two propositions in which we investigate the sets $\mathcal{M}_{L,Q}(0^+)$ for distance-based loss functions.

Proposition 4.13 *Let L be a convex, distance-based loss function whose associated ψ satisfies $\lim_{r\to\pm\infty}\psi(r) = \infty$, and let $Q \in \mathcal{Q}_{\mathbb{R}}$ with $\mathcal{C}_{L,Q}(t) < \infty$ for all $t \in \mathbb{R}$. Then $\mathcal{M}_{L,Q}(0^+)$ is a compact, non-empty interval. Moreover, if ψ is strictly convex then $\mathcal{M}_{L,Q}(0^+)$ contains exactly one element.*

Proof: Our first goal is to show that $\lim_{t\to\pm\infty}\mathcal{C}_{L,Q}(t) = \infty$. To this end let $(t_n) \subset \mathbb{R}$ be a sequence with $t_n \to -\infty$, and $B > 0$. Since $\lim_{r\to\pm\infty}\psi(r) = \infty$ there then exists an $r_0 > 0$ such that $\psi(r) \geq 2B$ for all $r \in \mathbb{R}$ with $|r| \geq r_0$. Since $Q(\mathbb{R}) = 1$ there is also an $M > 0$ with $Q([-M,M]) \geq 1/2$. Finally, there exists an $n_0 \geq 1$ with $t_n \leq -M - r_0$ for all $n \geq n_0$. For $y \in [-M,M]$ this yields $y - t_n \geq r_0$, and hence we find $\psi(y-t_n) \geq 2B$ for all $n \geq n_0$. From this we easily find

$$\mathcal{C}_{L,Q}(t_n) \geq \int_{[-M,M]} \psi(y-t_n)\,dQ(y) \geq 2B\,Q\big([-M,M]\big) = B\,,$$

i.e. we have shown $\mathcal{C}_{L,Q}(t_n) \to \infty$. Analogously we can show $\lim_{t\to\infty}\mathcal{C}_{L,Q}(t) = \infty$, and consequently we have $\lim_{t\to\pm\infty}\mathcal{C}_{L,Q}(t) = \infty$. Furthermore, the convexity of ψ implies that $t \mapsto \mathcal{C}_{L,Q}(t)$ is convex and hence it is continuous by the assumption $\mathcal{C}_{L,Q}(t) < \infty$, $t \in \mathbb{R}$. Now the assertions follow. ∎

Note that for distributions $Q \in \mathcal{Q}_{\text{bounded}}$ we automatically have $\mathcal{C}_{L,Q}(t) < \infty$ for all $t \in \mathbb{R}$. Furthermore, if L is of some growth type p then we have $\mathcal{C}_{L,Q}(t) < \infty$ for all $t \in \mathbb{R}$ and all $Q \in \mathcal{Q}_{\mathbb{R}}(L)$. Consequently, the above proposition gives $\mathcal{M}_{L,Q}(0^+) \neq \emptyset$ in both cases.

The following proposition compares $\mathcal{M}_{L,Q}(0^+)$ with the mean $\mathbb{E}Q$. Note that a similar result was independently found by A. Caponnetto in [6].

Proposition 4.14 *Let $M > 0$, and L be a distance-based, locally Lipschitz continuous loss function with associated ψ. Then we have:*

i) If $\mathbb{E}Q \in \mathcal{M}_{L,Q}(0^+)$ for all $Q \in \mathcal{Q}_{[-M,M],\text{sym}}$, then L is symmetric.

ii) If $\mathbb{E}Q \in \mathcal{M}_{L,Q}(0^+)$ *for all* $Q \in \mathcal{Q}_{\text{bounded}}$*, then there exists a constant* $c \geq 0$ *with* $L = c\, L_{\text{lsquares}}$*.*

Proof: Recall that the fundamental theorem of calculus for Lebesgue integrals shows that the derivative ψ' is (Lebesgue)-almost surely defined and integrable on every bounded interval.
i). Let us fix a $y \in [-M, M]$ such that ψ is differentiable at y and $-y$. We define $Q := \frac{1}{2}\delta_{\{-y\}} + \frac{1}{2}\delta_{\{y\}}$. Then we have $Q \in \mathcal{Q}_{\text{sym}}$ with $\mathbb{E}Q = 0$, and $\mathcal{C}_{L,Q}(t) = \frac{1}{2}\psi(-y-t) + \frac{1}{2}\psi(y-t)$. Therefore the derivative of $\mathcal{C}_{L,Q}(.)$ exists at 0 and can be computed by $\mathcal{C}'_{L,Q}(0) = -\frac{1}{2}\psi'(-y) - \frac{1}{2}\psi'(y)$. Furthermore, our assumption shows that $\mathcal{C}_{L,Q}(.)$ has a minimum at 0, and hence we have $0 = \mathcal{C}'_{L,Q}(0)$, i.e. $\psi'(-y) = -\psi'(y)$. According to our preliminary remark, this equality holds for almost all y, and thus the fundamental theorem of calculus for Lebesgue integrals shows that for all $y_0 \in \mathbb{R}$ we have

$$\psi(y_0) = \psi(0) + \int_0^{y_0} \psi'(t)dt = \psi(0) - \int_0^{y_0} \psi'(-t)dt = \psi(0) - \int_{-y_0}^{0} \psi'(t)dt = \psi(-y_0)\,.$$

ii). Let $y \neq 0$ and $\alpha > 0$ be real numbers such that ψ is differentiable at y, $-y$, and αy. We define $Q := \frac{\alpha}{1+\alpha}\delta_{\{0\}} + \frac{1}{1+\alpha}\delta_{\{(1+\alpha)y\}}$, so that we obtain $\mathbb{E}Q = y$ and $\mathcal{C}_{L,Q}(t) = \frac{\alpha}{1+\alpha}\psi(-t) + \frac{1}{1+\alpha}\psi(y+\alpha y-t)$ for all $t \in \mathbb{R}$. This shows that the derivative of $\mathcal{C}_{L,Q}(.)$ exists at y and can be computed by

$$\mathcal{C}'_{L,Q}(y) = -\frac{\alpha}{1+\alpha}\psi'(-y) - \frac{1}{1+\alpha}\psi'(\alpha y) = \frac{\alpha}{1+\alpha}\psi'(y) - \frac{1}{1+\alpha}\psi'(\alpha y)\,,$$

where in the last step we used *i)*. Again, our assumption $y \in \mathcal{M}_{L,Q}(0^+)$ gives $\mathcal{C}'_{L,Q}(y) = 0$ and hence we find $\alpha\psi'(y) = \psi'(\alpha y)$. Obviously, the latter holds for almost all $\alpha > 0$ and thus we obtain

$$\psi(ty) = \psi(0) + \int_0^t \psi'(sy)y\, ds = \int_0^t s\psi'(y)y\, ds = \frac{\psi'(y)}{2y}(ty)^2$$

for all $t > 0$. From this we easily obtain the assertion. ∎

The above proposition shows that there are only trivial distance-based surrogates for the least squares loss if one is interested in the entire class $\mathcal{Q}(L_{\text{lsquares}}) = \{Q \in \mathcal{Q}_{\mathbb{R}} : \mathbb{E}_{y\sim Q} y^2 < \infty\}$. Furthermore, it shows that the least squares loss is essentially the only distance-based loss function whose minimizer is the mean for all distributions in $\mathcal{Q}(L_{\text{lsquares}})$, i.e. if we are actually interested in finding the regression function $x \mapsto \mathbb{E}_P(Y|x)$, and we just know $\mathbb{E}_{x\sim P_X}\mathbb{E}_{y\sim P(.|x)} y^2 < \infty$, then the least squares loss is the only *distance-based loss* for this task[3] However, if we cannot ensure this tail assumption but know that the conditional distributions $P(.|x)$ are *symmetric* then the above proposition suggests that we may actually have alternatives to the least squares loss. In order to investigate this conjecture systematically we first need a target loss function that describes the goal of finding the mean. To this end let us consider the template loss $L_{\text{mean}} : \mathcal{Q}_{\mathbb{R}} \times \mathbb{R} \to [0, \infty)$ defined by

$$L_{\text{mean}}(Q, t) := |\mathbb{E}Q - t|\,, \qquad t \in \mathbb{R}\,, Q \in \mathcal{Q}_{\mathbb{R}}\,.$$

Note that this indeed defines a template loss since given a distribution P on $X \times \mathbb{R}$ of $\mathcal{Q}_{\mathbb{R}}$-type it is easy to see that $(x, t) \mapsto L_{\text{mean}}(P(.|x), t) = |\mathbb{E}_P(Y|x) - t|$ is measurable. Moreover we have

$$L^2_{\text{mean}}(Q, t) = \big(\mathbb{E}Q - t\big)^2 = \mathcal{C}_{L_{\text{lsquares}},Q}(t) - \mathcal{C}^*_{L_{\text{lsquares}},Q}$$

[3]The result in [6] shows that any *convex* loss function $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ satisfying $\mathbb{E}Q \in \mathcal{M}_{L,Q}(0^+)$ for all $Q \in \mathcal{Q}_{\text{bounded}}$ must be of the form $L(y, t) = c \cdot (y - t)^2 + h(y)$, $(y, t) \in \mathbb{R}^2$, where $h : \mathbb{R} \to \mathbb{R}$ is a suitable function. Consequently, other loss functions (e.g. Bregman divergences) for estimating the mean cannot be convex and hence may lead to algorithmic problems. Whether such problems can be resolved by estimating a suitable transform of the mean is, as far as we know, open, and because of space constraints we do not further investigate this highly interesting question.

for all $Q \in \mathcal{Q}(L_{\text{lsquares}})$ and $t \in \mathbb{R}$. Since the minimal L_{mean}-risks equal 0, i.e. $\mathcal{C}^*_{L_{\text{mean}},Q} = 0$, we consequently have $\mathcal{M}_{L_{\text{mean}},Q}(\sqrt{\varepsilon}) = \mathcal{M}_{L_{\text{lsquares}},Q}(\varepsilon)$ for all $\varepsilon > 0$, and from this we obtain

$$\delta_{\max,L_{\text{mean}},L}(\sqrt{\varepsilon}, Q) = \delta_{\max,L_{\text{lsquares}},L}(\varepsilon, Q) \tag{40}$$

for all distance-based losses L, all $Q \in \mathcal{Q}(L_{\text{lsquares}}) \cap \mathcal{Q}(L)$, and all $\varepsilon \in [0, \infty]$. In other words, by considering L_{mean}-calibration we simultaneously obtain results on L_{lsquares}-calibration. Furthermore, if $t \mapsto \mathcal{C}_{L,Q}(t)$ has a *unique* minimum at $\mathbb{E}Q$ then we obviously have $L_{\text{mean}}(Q,t) = \breve{L}(Q,t)$ for all $t \in \mathbb{R}$, and consequently we obtain

$$\delta_{\max,L_{\text{mean}},L}(\varepsilon, Q) = \delta_{\max,\breve{L},L}(\varepsilon, Q)\,, \qquad \varepsilon \in [0, \infty]. \tag{41}$$

In other words, by considering L_{mean}-calibration of L we will also gain some insight into the self-calibration properties of L.

Now, the following key lemma presents an alternative way to compute the inner risks $\mathcal{C}_{L,Q}(.)$ when both L and Q are symmetric:

Lemma 4.15 *Let L be a symmetric loss function with associated ψ, and $Q \in \mathcal{Q}_{\text{sym}}(L)$. For $t \in \mathbb{R}$ we then have*

$$\mathcal{C}_{L,Q}(\mathbb{E}Q + t) = \mathcal{C}_{L,Q}(\mathbb{E}Q - t) = \frac{1}{2}\int_{\mathbb{R}} \big(\psi(y - \mathbb{E}Q - t) + \psi(y - \mathbb{E}Q + t)\big)\, dQ(y). \tag{42}$$

In addition, if L is convex we have $\mathcal{C}_{L,Q}(\mathbb{E}Q) = \mathcal{C}^*_{L,Q}$, and if L is actually strictly convex we also have $\mathcal{C}_{L,Q}(\mathbb{E}Q + t) > \mathcal{C}^*_{L,Q}$ for all $t \neq 0$.

Proof: Let us fix a $Q \in \mathcal{Q}_{\text{sym}}(L)$ and write $m := \mathbb{E}Q$. The symmetry of $Q^{(m)}$ and ψ then yields

$$\mathcal{C}_{L,Q}(m+t) = \int_{\mathbb{R}} \psi(y-t) dQ^{(m)}(y) = \int_{\mathbb{R}} \psi(-y-t) dQ^{(m)}(y) = \int_{\mathbb{R}} \psi(y+t) dQ^{(m)}(y) = \mathcal{C}_{L,Q}(m-t).$$

Since this gives $\mathcal{C}_{L,Q}(m+t) = \frac{1}{2}\big(\mathcal{C}_{L,Q}(m+t) + \mathcal{C}_{L,Q}(m-t)\big)$ we also obtain the second equation. Furthermore, if ψ is convex we can easily conclude

$$\mathcal{C}_{L,Q}(m+t) = \frac{1}{2}\int_{\mathbb{R}} \big(\psi(y-t) + \psi(y+t)\big) dQ^{(m)}(y) \geq \int_{\mathbb{R}} \psi(y) dQ^{(m)}(y) = \mathcal{C}_{L,Q}(m)$$

for all $t \in \mathbb{R}$. This shows the second assertion. The third assertion can be shown analogously. ■

With the help of the above lemma we can derive a simple formula for the calibration function $\delta_{\max,L_{\text{mean}},L}(\varepsilon, Q)$ if L is convex:

Lemma 4.16 *Let L be a symmetric, convex loss function and $Q \in \mathcal{Q}_{\text{sym}}(L)$. Then we have*

$$\delta_{\max,L_{\text{mean}},L}(\varepsilon, Q) = \mathcal{C}_{L,Q}(\mathbb{E}Q + \varepsilon) - \mathcal{C}^*_{L,Q}\,, \qquad \varepsilon \geq 0. \tag{43}$$

In particular, $\varepsilon \mapsto \delta_{\max,L_{\text{mean}},L}(\varepsilon, Q)$ is convex.

Proof: Obviously, we have $\mathcal{M}_{L_{\text{mean}},Q}(\varepsilon) = (\mathbb{E}Q - \varepsilon, \mathbb{E}Q + \varepsilon)$. Since this set is open, it is easy to see that the continuity assumption in Lemma 4.15 is superfluous and hence the assertion follows by Lemma 2.11 and Lemma 4.15. ■

Our next result is a rather technical lemma which will be used at various times to establish *upper* bounds on $\delta_{\max}(\varepsilon, Q)$. For its formulation we need the set

$$\mathcal{Q}^*_{\mathrm{sym}} := \big\{Q \in \mathcal{Q}_{\mathrm{sym}} : Q\big(\mathbb{E}Q + (-\rho, \rho)\big) > 0 \text{ for all } \rho > 0\big\}$$

which contains all symmetric distributions that do not vanish around their means. Moreover, we also need the sets $\mathcal{Q}^*_{I,\mathrm{sym}} := \mathcal{Q}_I \cap \mathcal{Q}^*_{\mathrm{sym}}$ for I being a bounded interval, and $\mathcal{Q}^*_{\mathrm{bounded,sym}} := \mathcal{Q}_{\mathrm{bounded}} \cap \mathcal{Q}^*_{\mathrm{sym}}$. Now the result reads as follows:

Lemma 4.17 *Let L be a symmetric, continuous loss function with associated ψ. Assume that there exist $\delta_0 \in \mathbb{R}$, $s_1, s_2 \in \mathbb{R}$, $s_1 \neq s_2$ and an $\varepsilon_0 > 0$ such that for all $\varepsilon \in [0, \varepsilon_0]$ we have*

$$\frac{\psi(s_1 + \varepsilon) + \psi(s_2 + \varepsilon)}{2} - \psi\Big(\frac{s_1 + s_2}{2} + \varepsilon\Big) \leq \delta_0 \,. \tag{44}$$

Then for $M := |\frac{s_1+s_2}{2}| + \varepsilon_0$ and all $\delta > 0$ there exists a Lebesgue-absolutely continuous $Q \in \mathcal{Q}^*_{[-M,M],\mathrm{sym}}(L)$ with $\mathbb{E}Q = 0$ such that for $t := \frac{s_2 - s_1}{2}$ we have

$$\mathcal{C}_{L,Q}(t) - \mathcal{C}_{L,Q}(0) \leq \delta_0 + \delta \,.$$

Furthermore, there exists a Lebesgue-absolutely continuous $Q \in \mathcal{Q}_{[-M,M],\mathrm{sym}}(L)$ with $\mathbb{E}Q = 0$ and

$$\mathcal{C}_{L,Q}(t) - \mathcal{C}_{L,Q}(0) \leq \delta_0 \,.$$

Proof: We write $y_0 := \frac{s_1+s_2}{2}$. Furthermore, if $y_0 = 0$ we let Q be the uniform distribution $\mu_{[-\varepsilon_0,\varepsilon_0]}$ on $[-\varepsilon_0, \varepsilon_0]$. Otherwise, we define

$$Q := \alpha\mu_{[-|\frac{y_0}{2}|,|\frac{y_0}{2}|]} + \frac{1-\alpha}{2}\mu_{[-y_0-\varepsilon_0,-y_0]} + \frac{1-\alpha}{2}\mu_{[y_0,y_0+\varepsilon_0]} \,,$$

where $\alpha \in (0,1)$ is a real number satisfying

$$\sup_{y \in [-|\frac{y_0}{2}|,|\frac{y_0}{2}|]} \Big|\frac{\psi(y-t) + \psi(y+t)}{2} - \psi(y)\Big| \leq \frac{\delta}{\alpha} \,.$$

Now, in the case $y_0 \neq 0$ this construction yields $\mathbb{E}Q = 0$ and

$$\begin{aligned}
\mathcal{C}_{L,Q}(t) - \mathcal{C}_{L,Q}(0) &= \int_{\mathbb{R}} \frac{\psi(y-t) + \psi(y+t)}{2} - \psi(y) dQ(y) \\
&= \alpha \int_{[-|\frac{y_0}{2}|,|\frac{y_0}{2}|]} \frac{\psi(y-t) + \psi(y+t)}{2} - \psi(y)\, d\mu_{[-|\frac{y_0}{2}|,|\frac{y_0}{2}|]}(y) \\
&\quad + (1-\alpha) \int_{[y_0,y_0+\varepsilon_0]} \frac{\psi(y-t) + \psi(y+t)}{2} - \psi(y)\, d\mu_{[y_0,y_0+\varepsilon_0]}(y) \\
&\leq \delta + (1-\alpha) \int_{[0,\varepsilon_0]} \frac{\psi(s_1+\varepsilon) + \psi(s_2+\varepsilon)}{2} - \psi\Big(\frac{s_1+s_2}{2} + \varepsilon\Big)\, d\mu_{[0,\varepsilon_0]}(\varepsilon) \\
&\leq \delta_0 + \delta \,.
\end{aligned}$$

Furthermore, the case $y_0 = 0$ can be shown analogously, and the last assertion follows if we repeat the above construction with $\alpha = 0$. ∎

With the above preparations we can finally establish our first main result that characterizes loss functions L that are L_{mean}-calibrated with respect to $\mathcal{Q}^*_{\mathrm{sym}}(L)$:

Theorem 4.18 *Let L be a symmetric, continuous loss function. Then the following statements are equivalent:*

i) L is L_{mean}-calibrated with respect to $\mathcal{Q}^*_{\text{sym}}(L)$.

ii) L is convex and its associated function ψ has its only minimum at 0.

Proof: *ii)* $\Rightarrow$ *i)*. Assume that L is not L_{mean}-calibrated with respect to $\mathcal{Q}^*_{\text{sym}}(L)$. By Lemma 4.16 there then exist a $Q \in \mathcal{Q}^*_{\text{sym}}(L)$ and a $t \neq 0$ with $\mathcal{C}_{L,Q}(m+t) = \mathcal{C}^*_{L,Q}$, where $m := \mathbb{E}Q$. Using $\mathcal{C}_{L,Q}(m) = \mathcal{C}^*_{L,Q}$, which we know from Lemma 4.15, then yields

$$\int_{\mathbb{R}} \frac{\psi(y-t)+\psi(y+t)}{2} - \psi(y)\, dQ^{(m)}(y) = \mathcal{C}_{L,Q}(m+t) - \mathcal{C}_{L,Q}(m) = 0\,,$$

and hence the convexity of ψ shows $\frac{\psi(y-m-t)+\psi(y-m+t)}{2} - \psi(y-m) = 0$ for Q-almost all $y \in \mathbb{R}$. The continuity of ψ and the assumption $Q\big(m+(-\rho,\rho)\big) > 0$ for all $\rho > 0$, then guarantee that $\frac{\psi(y-m-t)+\psi(y-m+t)}{2} - \psi(y-m) = 0$ holds for $y := m$. However, by the symmetry of ψ this implies $\psi(t) = \psi(0)$ which violates our assumption on ψ.
i) $\Rightarrow$ *ii)*. Assume that ψ is not convex. Then Lemma A.1 shows that there exist $s_1, s_2 \in \mathbb{R}$ with $\frac{\psi(s_1)+\psi(s_2)}{2} - \psi(\frac{s_1+s_2}{2}) < 0$. With the continuity of ψ we then find (44) for some suitable $\delta_0 < 0$ and $\varepsilon_0 > 0$, and consequently Lemma 4.17 gives a $Q \in \mathcal{Q}^*_{[-M,M],\text{sym}}(L)$ and a $t \neq 0$ with $\mathbb{E}Q = 0$ and $\mathcal{C}_{L,Q}(t) < \mathcal{C}_{L,Q}(0)$. Now observe that since ψ is continuous and Q has bounded support, the map $t \mapsto \mathcal{C}_{L,Q}(t)$ is continuous on $\mathbb{R}$. Let $(t_n) \subset \mathbb{R}$ be a sequence with $\mathcal{C}_{L,Q}(t_n) \to \mathcal{C}^*_{L,Q}$ for $n \to \infty$. Now, our previous considerations showed $\mathcal{C}_{L,Q}(0) \neq \mathcal{C}^*_{L,Q}$ and hence (t_n) is eventually bounded away from 0, i.e. there exist an $\varepsilon > 0$ and an $n_0 \in \mathbb{N}$ such that $|t_n| \geq \varepsilon$ for all $n \geq n_0$. This gives

$$\delta_{\max}(\varepsilon, Q) = \inf_{t' \notin (-\varepsilon,\varepsilon)} \mathcal{C}_{L,Q}(t) - \mathcal{C}^*_{L,Q} \;\leq\; \mathcal{C}_{L,Q}(t_n) - \mathcal{C}^*_{L,Q}\,,$$

for all $n \geq n_0$. For $n \to \infty$ we hence find $\delta_{\max}(\varepsilon, Q) = 0$, and consequently L is convex. Finally, assume that there exists a $t \neq 0$ with $\psi(t) = \psi(0)$. Then we find $\mathcal{C}_{L,Q}(t) = \mathcal{C}^*_{L,Q}$ for the distribution Q defined by $Q(\{0\}) = 1$, and hence we obtain $\delta_{\max}(t, Q) = 0$ by Lemma 4.16. ∎

The following theorem considers calibration with respect to the larger class $\mathcal{Q}_{\text{sym}}$:

Theorem 4.19 *Let L be a symmetric, continuous loss function. Then the following statements are equivalent:*

i) L is L_{mean}-calibrated with respect to $\mathcal{Q}_{\text{sym}}(L)$.

ii) L is strictly convex.

Proof: If L is strictly convex then Lemma 4.15 and Lemma 4.16 show that L is L_{mean}-calibrated with respect to $\mathcal{Q}_{\text{sym}}(L)$. Conversely, if L is L_{mean}-calibrated with respect to $\mathcal{Q}_{\text{sym}}(L)$, then Theorem 4.18 shows that L is convex. Let us suppose that its associated $\psi : \mathbb{R} \to [0,\infty)$ is not strictly convex. Then there exists $r_1, r_2 \in \mathbb{R}$ with $r_1 \neq r_2$ and

$$\psi\Big(\frac{1}{2}r_1 + \frac{1}{2}r_2\Big) = \frac{1}{2}\psi(r_1) + \frac{1}{2}\psi(r_2)\,.$$

From this and Lemma A.1 we easily find (44) for $\delta_0 = 0$ and some suitable $s_1 \neq s_2$ and $\varepsilon_0 > 0$. Lemma 4.17 then gives a $t_0 \neq 0$ and a $Q \in \mathcal{Q}_{[-M,M],\text{sym}}(L)$ with $\mathcal{C}_{L,Q}(\mathbb{E}Q + t_0) = \mathcal{C}^*_{L,Q}$ and hence Lemma 4.16 shows that L is not L_{mean}-calibrated with respect to $\mathcal{Q}_{\text{sym}}(L)$. ∎

Our next aim is to estimate the function $\varepsilon \mapsto \delta_{\max}(\varepsilon, \mathcal{Q})$ for some classes $\mathcal{Q} \subset \mathcal{Q}_{\mathrm{sym}}$. To this end we define the *modulus of convexity* of a function $f : I \to \mathbb{R}$, where I is an interval, by

$$\delta_f(\varepsilon) := \inf\left\{ \frac{f(x_1) + f(x_2)}{2} - f\Big(\frac{x_1 + x_2}{2}\Big) : x_1, x_2 \in I \text{ with } |x_1 - x_2| \geq \varepsilon \right\}, \qquad \varepsilon > 0.$$

In addition we say that f is *uniformly convex* if $\delta_f(\varepsilon) > 0$ for all $\varepsilon > 0$. Some properties of the modulus of convexity and uniformly convex functions can be found in Appendix.

With the help of the modulus of convexity we can now formulate the following theorem that estimates $\delta_{\max}(\varepsilon, \mathcal{Q})$ and characterizes uniform calibration:

Theorem 4.20 *Let L be a symmetric, convex loss function with associated ψ. Then we have:*

i) For all $M > 0$, $\varepsilon > 0$, and all $\mathcal{Q}$ with $\mathcal{Q}^*_{[-M,M],\mathrm{sym}} \subset \mathcal{Q} \subset \mathcal{Q}_{[-M,M],\mathrm{sym}}$ we have

$$\delta_{\psi_{|[-(2M+\varepsilon),2M+\varepsilon]}}(2\varepsilon) \;\leq\; \delta_{\max}(\varepsilon, \mathcal{Q}) \;\leq\; \delta_{\psi_{|[-M/2,M/2]}}(2\varepsilon)\,. \tag{45}$$

Consequently, the following statements are equivalent:

(a) L is uniformly L_{mean}-calibrated with respect to $\mathcal{Q}^*_{[-M,M],\mathrm{sym}}$ for all $M > 0$.

(b) L is uniformly L_{mean}-calibrated with respect to $\mathcal{Q}_{[-M,M],\mathrm{sym}}$ for all $M > 0$.

(c) The function ψ is strictly convex.

ii) For all $\varepsilon > 0$ we have

$$\delta_\psi(2\varepsilon) = \delta_{\max}(\varepsilon, \mathcal{Q}_{\mathrm{sym}}(L)) = \delta_{\max}(\varepsilon, \mathcal{Q}^*_{\mathrm{bounded,sym}}(L))\,. \tag{46}$$

Consequently, the following statements are equivalent:

(a) L is uniformly L_{mean}-calibrated with respect to $\mathcal{Q}_{\mathrm{sym}}(L)$.

(b) L is uniformly L_{mean}-calibrated with respect to $\mathcal{Q}^*_{\mathrm{bounded,sym}}(L)$.

(c) The function ψ is uniformly convex.

Proof: *i).* For $Q \in \mathcal{Q}_{[-M,M],\mathrm{sym}}$ we have $\mathbb{E}Q \in [-M, M]$, and hence we find

$$\delta_{\max}(\varepsilon, Q) = \int_{[-M,M]} \frac{\psi\big(y - \mathbb{E}Q - \varepsilon\big) + \psi\big(y - \mathbb{E}Q + \varepsilon\big)}{2} - \psi(y - \mathbb{E}Q)\, dQ(y) \geq \delta_{\psi_{|[-(2M+\varepsilon),2M+\varepsilon]}}(2\varepsilon)$$

by Lemma 4.15 and Lemma 4.16. This shows the first inequality. To prove the second inequality let $n \geq 1$. Then there exist $s_1, s_2 \in [-M/2, M/2]$ with $s_1 - s_2 \geq 2\varepsilon$ and

$$\frac{\psi(s_1) + \psi(s_2)}{2} - \psi\left(\frac{s_1 + s_2}{2}\right) < \delta_{\psi_{|[-M/2,M/2]}}(2\varepsilon) + \frac{1}{n} =: \delta_0\,.$$

By the continuity of ψ there thus exists an $\varepsilon_0 \in (0, M]$ such that (44) is satisfied for δ_0, and consequently Lemma 4.17 gives a $Q \in \mathcal{Q}^*_{[-M,M],\mathrm{sym}}$ with $\mathbb{E}Q = 0$ and with

$$\mathcal{C}_{L,Q}(t) - \mathcal{C}^*_{L,Q} \leq \delta_{\psi_{|[-M/2,M/2]}}(2\varepsilon) + \frac{2}{n}$$

for $t := \frac{s_1 - s_2}{2}$. Since $t \geq \varepsilon$ we also have $\delta_{\max}(\varepsilon, Q) \leq \mathcal{C}_{L,Q}(t) - \mathcal{C}^*_{L,Q}$, and hence we find

$$\delta_{\max}(\varepsilon, \mathcal{Q}^*_{[-M,M],\mathrm{sym}}) \leq \delta_{\max}(\varepsilon, Q) \leq \delta_{\psi_{|[-M/2,M/2]}}(2\varepsilon) + \frac{2}{n}.$$

Loss function	$\psi(t)$	lower bound of $\delta_{\psi_{\|[-B,B]}}$	upper bound of $\delta_{\psi_{\|[-B,B]}}$
L_1-loss	$\|t\|$	0	0
L_p-loss, $p \in (1,2)$	$\|t\|^p$	$\frac{p(p-1)}{8} B^{p-2}\varepsilon^2$	$\frac{p}{8(p-1)^2} B^{p-2}\varepsilon^2$
L_p-loss, $p \in [2,\infty)$	$\|t\|^p$	$\left(\frac{\varepsilon}{2}\right)^p$	$\left(\frac{\varepsilon}{2}\right)^p$
Logistic loss	$-\ln \frac{4e^t}{(1+e^t)^2}$	$\frac{e^{\varepsilon/2}-1}{2e^{\varepsilon/2}} \ln \frac{e^B+e^\varepsilon}{e^B+e^{\varepsilon/2}}$	$\frac{e^{\varepsilon/2}-1}{e^{\varepsilon/2}} \ln \frac{e^B+e^\varepsilon}{e^B+e^{\varepsilon/2}}$
Huber's loss, $c>0$	$\frac{t^2}{2}$ if $\|t\|\le c$ $c\|t\|-\frac{c^2}{2}$ else	$\frac{\varepsilon^2}{8}$ if $B \le c$ 0 else	$\frac{\varepsilon^2}{8}$ if $B \le c$ 0 else

Table 3: Some invariant loss functions and corresponding upper and lower bounds of $\delta_{\psi_{\|[-B,B]}}(\varepsilon)$, $0 < \varepsilon \le B$, for the restriction $\psi_{\|[-B,B]}$ of ψ to $[-B,B]$, for $B > 0$. The asymptotic behaviour for the L_p-loss, $1 < p < 2$, is computed in Example A.4. For the L_p-loss, $p \ge 2$, and Huber's loss the lower bounds can be found by Clarkson's inequality, and the upper bounds were found by finding suitable $x_1, x_2 \in [B,B]$. The calculations for the logistic loss can be found in Example A.5.

Since this holds for all $n \ge 1$ the second inequality follows. Finally, Lemma A.1 shows that ψ is strictly convex if and only if $\delta_{\psi_{\|[-B,B]}}(\varepsilon) > 0$ for all $B, \varepsilon > 0$, and hence the characterization follows. *ii).* Analogously to the proof of the first inequality in (45) we find $\delta_\psi(2\varepsilon) \le \delta_{\max}(\varepsilon, \mathcal{Q}_{\text{sym}}(L))$ for all $\varepsilon > 0$. Furthermore by the second inequality in (45) we obtain $\delta_{\max}(\varepsilon, \mathcal{Q}^*_{\text{bounded,sym}}(L)) \le \delta_\psi(2\varepsilon)$ for all $\varepsilon > 0$, and hence (46) is proved. Finally, the characterization is a consequence of (46). ■

Remark 4.21 The above theorem shows that the modulus of convexity completely determines whether a loss function is uniformly L_{mean}-calibrated with respect $\mathcal{Q}_{\text{sym}}(L)$ or $\mathcal{Q}^*_{\text{bounded,sym}}(L)$. Unfortunately, Lemma A.3 shows that for all distance-based loss functions of growth type $p < 2$ we have $\delta_\psi(\varepsilon) = 0$ for *all* $\varepsilon > 0$. In particular, Lipschitz continuous, distance-based losses which are of special interest for robust regression methods (see e.g. [8]) are not uniformly calibrated with respect to $\mathcal{Q}_{\text{sym}}(L)$ or $\mathcal{Q}^*_{\text{bounded,sym}}(L)$, and consequently we cannot establish *strong* relations between the excess L-risks and $\mathcal{R}_{L_{\text{mean}},P}(.)$ in the sense of Question 2.

On the other hand, note that symmetric, strictly convex losses L are L_{mean}-calibrated with respect to $\mathcal{Q}_{\text{sym}}(L)$, and hence we can show analogously to Theorem 3.16 that $f_n \to \mathbb{E}(Y|.)$ in probability, whenever $\mathcal{R}_{L,P}(f_n) \to \mathcal{R}^*_{L,P}$ and P is of type $\mathcal{Q}_{\text{sym}}(L)$. In addition, if we restrict our considerations to $\mathcal{Q}_{[-M,M],\text{sym}}$ or $\mathcal{Q}^*_{[-M,M],\text{sym}}$ then every strictly convex loss becomes uniformly L_{mean}-calibrated, and in this case $\delta_{\psi_{\|[-B,B]}}(.)$, $B > 0$, can be used to describe the corresponding calibration function. For some important losses we have listed the behaviour of $\delta_{\psi_{\|[-B,B]}}(.)$ in Table 3. Furthermore, Lemma A.3 establishes a formula for the modulus of convexity which often helps to bound the modulus.

In Theorem 4.18 we have seen that for $Q \in \mathcal{Q}^*_{\text{sym}}(L)$ we may have $\delta_{\max}(\varepsilon, Q) > 0$, $\varepsilon > 0$, even if L is not strictly convex. The key reason for this possibility was the assumption that Q has some mass around its center. Now recall that in the proof of the upper bounds of Theorem 4.20 we used the fact that for general $Q \in \mathcal{Q}^*_{\text{sym}}$ this mass can be arbitrarily small. However, if we enforce lower bounds on this mass the construction of this proof no longer works. Instead, it turns out that we can establish lower bounds on $\delta_{\max}(\varepsilon, Q)$ as the following example illustrates:

Example 4.22 Let L be the distance-based loss function whose associated ψ is $\psi(t) = |t|$, $t \in \mathbb{R}$. Then for all $Q \in \mathcal{Q}_{\text{sym}}(L)$, $m := \mathbb{E}Q$, and all $\varepsilon > 0$ we have

$$\delta_{\max}(\varepsilon, Q) = \int_0^\varepsilon Q^{(m)}\big((-s,s)\big)\, ds\,.$$

To see this we first observe that for $t \geq 0$ and $y \in \mathbb{R}$ an easy calculation shows

$$\psi(y-t) - \psi(y) = \begin{cases} t & \text{if } y \leq 0 \\ t - 2y & \text{if } y \in (0,t) \\ -t & \text{if } y \geq t\,. \end{cases}$$

Consequently the symmetry of $Q^{(m)}$ yields

$$\begin{aligned} \mathcal{C}_{L,Q}(m+t) - \mathcal{C}^*_{L,Q} &= tQ^{(m)}\big((-\infty,0]\big) + tQ^{(m)}\big((0,t)\big) - 2\int_{0<y<t} y\,dQ^{(m)}(y) - tQ^{(m)}\big([t,\infty)\big) \\ &= tQ^{(m)}\big((-t,t)\big) - 2\int_{0\leq y<t} y\,dQ^{(m)}(y) \\ &= \int_0^t Q^{(m)}\big((-t,t)\big) - Q^{(m)}\big((-t,-s]\cup[s,t)\big)\,ds \\ &= \int_0^t Q^{(m)}\big((-s,s)\big)\,ds\,. \end{aligned}$$

From this we easily find the assertion by Lemma 4.16.

Remark 4.23 The above results show that using distance-based loss functions for regression problems requires some care: for example let us suppose that the primary goal of the regression problem is to find the regression function. If we only know that the noise distributions have finite variances (and expect that these distributions are rather asymmetric) then the least squares loss is the only reasonable distance-based choice by Proposition 4.14. However, if we know that the noise is (almost) symmetric then e.g. symmetric, strictly convex and Lipschitz continuous losses like the logistic loss can be a reasonable alternative. In addition, if we are confident that the noise is rather concentrated around its mean, e.g. in the form of $Q^{(m)}((-s,s)) > c_Q s^q$ for small $s > 0$, then even convex loss functions like the absolute distance loss considered in the previous example can be a good choice. Finally, if we additionally expect that the data set contains extreme outliers then the logistic loss or the absolute distance loss may actually be the better choice than the least squares loss. However, recall that such a decision only makes sense under an (almost) symmetric behaviour of the noise distribution.

When we introduced the template loss L_{mean} we also discussed its relation to self-calibration issues. Therefore let us finally investigate in which sense convex, distance-based loss functions are self-calibrated.

Theorem 4.24 *Let L be a distance-based, convex loss function whose associated ψ satisfies $\lim_{t\to\pm\infty}\psi(t) = \infty$. Then we have:*

i) L is self-calibrated with respect to $\mathcal{Q}_{\text{bounded}}$.

ii) If L is of some growth type $p \geq 1$ then L is self-calibrated with respect to $\mathcal{Q}_{\mathbb{R}}(L)$.

iii) If L is strictly convex then L is uniformly self-calibrated with respect to $\mathcal{Q}_{[-M,M]}$ for all $M > 0$, and we have

$$\delta_{\max,\check{L},L}(\varepsilon, \mathcal{Q}_{[-M,M]}) \geq 2\delta_{\psi_{|[-(2M+\varepsilon),2M+\varepsilon]}}(\varepsilon)\,, \qquad \varepsilon > 0,\ M > 0. \tag{47}$$

iv) If L is uniformly convex then L is uniformly self-calibrated with respect to $\mathcal{Q}_{\text{bounded}}$ and we have

$$\delta_{\max,\check{L},L}(\varepsilon, \mathcal{Q}_{\text{bounded}}) \geq 2\delta_{\psi}(\varepsilon)\,, \qquad \varepsilon > 0. \tag{48}$$

v) If L is uniformly convex and of some growth type $p \geq 2$ then L is uniformly self-calibrated with respect to $\mathcal{Q}_{\mathbb{R}}(L)$ and we have

$$\delta_{\max,\check{L},L}(\varepsilon, \mathcal{Q}_{\mathbb{R}}(L)) \geq 2\delta_{\psi}(\varepsilon)\,, \qquad \varepsilon > 0. \tag{49}$$

Proof: *i).* Proposition 4.13 shows $\mathcal{Q}_{\text{bounded}} \subset \mathcal{Q}_{\min}(L)$, and hence Lemma 3.15 implies *i).*
ii). The results in [8] together with Proposition 4.13 show $\mathcal{Q}_{\mathbb{R}}(L) \subset \mathcal{Q}_{\min}(L)$. Again, the assertion then follows from Lemma 3.15.
iii). Let us fix a $Q \in \mathcal{Q}_{[-M,M]}$. By Proposition 4.13 we then know that there exists a $t_Q^* \in \mathbb{R}$ with $\mathcal{M}_{L,Q}(0^+) = \{t_Q^*\}$, and consequently (10) reduces to

$$\delta_{\max,\check{L},L}(\varepsilon, Q) = \min\bigl\{\mathcal{C}_{L,Q}(t_Q^* - \varepsilon) - \mathcal{C}_{L,Q}^*, \mathcal{C}_{L,Q}(t_Q^* + \varepsilon) - \mathcal{C}_{L,Q}^*\bigr\}\,. \tag{50}$$

Let us now assume that $t_Q^* > M$. Then we have $y - t_Q^* \leq y - M \leq 0$ for all $y \in [-M, M]$. Since ψ is convex and has a minimum at 0 it is decreasing on $(-\infty, 0]$ and hence we find $\psi(y-M) \leq \psi(y-t_Q^*)$ for all $y \in [-M, M]$. This directly implies $\mathcal{C}_{L,Q}(M) \leq \mathcal{C}_{L,Q}(t_Q^*)$, and hence our assumption $t_Q^* > M$ cannot be true. Since we can analogously show $t_Q^* \geq -M$, we have $t_Q^* \in [-M, M]$. Moreover, for $t \in [-M - \varepsilon, M + \varepsilon]$ we have

$$\begin{aligned}
\frac{\mathcal{C}_{L,Q}(t) + \mathcal{C}_{L,Q}^*}{2} &= \int_{[-M,M]} \frac{\psi(y-t) + \psi(y-t_Q^*)}{2}\, dQ(y) \\
&\geq \int_{[-M,M]} \psi\Bigl(y - \frac{t + t_Q^*}{2}\Bigr) + \delta_{\psi_{|[-(2M+\varepsilon),2M+\varepsilon]}}\bigl(|t - t_Q^*|\bigr)\, dQ(y) \\
&= \mathcal{C}_{L,Q}\Bigl(\frac{t + t_Q^*}{2}\Bigr) + \delta_{\psi_{|[-(2M+\varepsilon),2M+\varepsilon]}}\bigl(|t - t_Q^*|\bigr) \\
&\geq \mathcal{C}_{L,Q}^* + \delta_{\psi_{|[-(2M+\varepsilon),2M+\varepsilon]}}\bigl(|t - t_Q^*|\bigr)\,,
\end{aligned} \tag{51}$$

and thus we find (47) by combining (50) with (51) and $t_Q^* \in [-M, M]$.
iv). Inequality (47) immediately implies (48).
v). The proof of Inequality (49) is analogous to the proof of (47). ∎

Remark 4.25 The above theorem shows that for convex, distance-based losses L, approximate L-risk minimizers approximate the Bayes decision functions in the sense of Theorem 3.16, i.e. in probability. In particular note that the absolute distance loss can be used to estimate the median (multi)-function in this weak sense. Moreover, Example 4.22 together with Theorem 2.17 shows that in order to estimate the median function in a stronger sense one needs assumptions on the concentration of the noise distributions. The reason for this observation is the fact that for *symmetric, convex* losses L whose associated ψ have a *unique* minimum at 0 the stronger convexity assumptions in *iii)-v)* are also *necessary* for uniform self-calibration. To see the latter observe that for such L Theorem 4.18 shows the L_{mean}-calibration with respect to $\mathcal{Q}_{\text{sym}}^*(L)$, and from Lemma 4.16 we may thus conclude that $\mathcal{C}_{L,Q}(.)$ has a *unique* minimum at $\mathbb{E}Q$ for all $Q \in \mathcal{Q}_{\text{sym}}^*(L)$. Equation (41) then shows

$$\delta_{\max,L_{\text{mean}},L}(\varepsilon, Q) = \delta_{\max,\check{L},L}(\varepsilon, Q)\,, \qquad \varepsilon \in [0,\infty]\,, Q \in \mathcal{Q}_{\text{sym}}^*(L). \tag{52}$$

Now assume e.g. that L is uniformly self-calibrated with respect to $\mathcal{Q}_{[-M,M]}$ for all $M > 0$. Then it is also uniformly self-calibrated with respect to $\mathcal{Q}_{[-M,M],\text{sym}}^*$, and thus (52) shows that L is L_{mean}-calibrated with respect to $\mathcal{Q}_{[-M,M],\text{sym}}^*$. By Theorem 4.20 we then see that L is strictly convex.

Remark 4.26 Inequalities for excess risks resulting from Theorem 4.24 can be used to establish *variance bounds* which are important for bounding the estimation error by Talagrand's inequality. Indeed, if e.g. L is a strictly convex loss function then its corresponding ψ is locally Lipschitz continuous and hence (32) can be used to find an "inner" version of a variance bound. This approach was somewhat implicitly taken in

e.g. [1, 4] in order to derive variance bounds for margin-based losses, but of course it also works for distance-based losses. However, it sometimes provides too loose bounds as e.g. the hinge loss shows. Indeed, this loss fails to be uniformly self-calibrated not only for $\eta \to 1/2$ but also for $\eta \to 0$ and $\eta \to 1$. In order to establish variance bounds using the self-calibration we consequently *need* not only to ensure a noise assumption in the sense of Tsybakov, but also a similar assumption ensuring that the set where η is either close to 0 or 1 is small. However, the assumption for η close to 0 and 1 is superfluous as the variance bound established in [29, Lemma 6.1] shows. One may ask whether variance bounds can also be *directly* established using our general framework and a (template) loss which reflects the variance bounds one is interested in. Some preliminary considerations we have already made in this directions are promising but due to space constraints we do not go into further details.

4.4 Density Level Detection

In this subsection we show how the developed theory can be used to investigate the density level detection problem. To this end let us first recall that in this *unsupervised* learning problem we have a *known* probability measure μ on X and the goal is to estimate a level set $\{g > \rho\}$ or $\{g \geq \rho\}$ of an *unknown* function $g : X \to [0, \infty)$ satisfying $\|g\|_{\mathcal{L}_1(\mu)} = 1$. The only information to achieve this goal is given to us by a data set $T := (x_1, \ldots, x_n)$ of n samples drawn independently from the probability measure $g\mu$. Typically, an estimate of a level set is of the form $\{f > 0\}$, where $f : X \to \mathbb{R}$ is a measurable function. In order to assess the quality of such an estimate one can use the loss function $L_{\text{DLD}} : X \times \mathbb{R} \to [0, \infty]$ defined by

$$L_{\text{DLD}}(x, t) := \mathbf{1}_{(-\infty,0)}\big((g(x) - \rho) \operatorname{sign} t\big)\,, \qquad x \in X,\, t \in \mathbb{R}\,. \tag{53}$$

Note that this loss function penalizes predictions t if either $t \geq 0$ and $g(x) < \rho$, or $t < 0$ and $g(x) > \rho$, whereas it completely ignores t if $g(x) = \rho$. In a slight abuse of notations we now write

$$\mathcal{R}_{L_{\text{DLD}},\mu}(f) := \int_X L_{\text{DLD}}\big(x, f(x)\big)\, d\mu(x)\,,$$

where $f : X \to \mathbb{R}$ is a measurable function. Obviously, this definition gives $\mathcal{R}_{L_{\text{DLD}},\mu}(f) = \mathcal{R}_{L_{\text{DLD}},P}(f)$ for every Polish space Y and every distribution P on $X \times Y$ with $P_X = \mu$. Moreover note, that if we additionally assume $\mu(\{g = \rho\}) = 0$, as it is usually done in the literature[4], we obtain the better known expression

$$\mathcal{R}_{L_{\text{DLD}},\mu}(f) = \mu\Big(\{g \geq \rho\} \triangle \{f \geq 0\}\Big)\,.$$

Now note that unlike for the supervised loss functions we *cannot compute* $L_{\text{DLD}}(x, t)$, since g is unknown to us, and consequently for training sets of the form $T = (x_1, \ldots, x_n) \in X^n$ we cannot use e.g. an ERM based on L_{DLD} simply because we cannot compute the empirical risk. To overcome this problem a framework was developed in [28] that translates the density level detection problem to a binary classification problem. The key idea of this framework was the following definition:

Definition 4.27 *Let μ be a distribution on X and $Y := \{-1, 1\}$. Furthermore, let $g : X \to [0, \infty]$ be a measurable function with $\|g\|_{\mathcal{L}_1(\mu)} = 1$. For $\rho > 0$ we define the distribution $P := g\mu \ominus_\rho \mu$ by*

$$\begin{aligned} P_X &:= \frac{g + \rho}{1 + \rho}\,\mu\,, \\ P(y = 1|x) &:= \frac{g(x)}{g(x) + \rho}\,, \qquad x \in X. \end{aligned}$$

[4] We could have gone this path, too. However, we will see that technically it is a bit more convenient to ignore the set $\{g = \rho\}$ rather than to assume that it (essentially) does not exist.

Note that the distribution $P := g\mu \ominus_\rho \mu$ describes a binary classification problem and consequently one could ask in which way the classification risk $\mathcal{R}_{L_{\text{class}},P}(.)$ is related to $\mathcal{R}_{L_{\text{DLD}},\mu}(.)$. It turned out in [28] that there is indeed a close relationship in between the two quantities. The goal in this subsection is to extend the considerations of [28] using our general theory. To this end we write $Y := \{-1, 1\}$ throughout this section. Furthermore, we always identify a distribution $Q \in \mathcal{Q}_Y$ by $\eta \in [0, 1]$ via the relationship $\eta := Q(\{1\})$. Now let us define $\bar{L}_{\text{DLD}} : [0, 1] \times \mathbb{R} \to [0, \infty]$ by

$$\bar{L}_{\text{DLD}}(\eta, t) := (1 - \eta)\mathbf{1}_{(-\infty,0)}\big((2\eta - 1)\operatorname{sign} t\big) . \tag{54}$$

Using the identification $\eta = Q(\{1\})$ we can then think of $\bar{L}_{\text{DLD}}$ as a *template* loss. Furthermore, for $P = g\mu \ominus_\rho \mu$ the P-instance $\bar{L}_{\text{DLD},P}$ of $\bar{L}_{\text{DLD}}$ is

$$\bar{L}_{\text{DLD},P}(x, t) = (1 - \eta(x))\mathbf{1}_{(-\infty,0)}\big((2\eta(x) - 1)\operatorname{sign} t\big) = \frac{\rho}{g(x) + \rho}\mathbf{1}_{(-\infty,0)}\big((g(x) - \rho)\operatorname{sign} t\big) ,$$

where $\eta(x) := P(y = 1|x) = \frac{g(x)}{g(x)+\rho}$. With this equation we then find

$$\mathcal{R}_{L_{\text{DLD}},\mu}(f) = \int_X L_{\text{DLD}}\big(x, f(x)\big)\frac{1 + \rho}{g(x) + \rho}\, dP_X(x) = \frac{1 + \rho}{\rho}\mathcal{R}_{\bar{L}_{\text{DLD},P},P}(f) \tag{55}$$

for all measurable $f : X \to \mathbb{R}$. Consequently, suitable supervised surrogates for the DLD-problem are exactly the losses that are (uniformly) $\bar{L}_{\text{DLD}}$-calibrated. In order to identify the latter losses let us first compute the corresponding calibration functions:

Lemma 4.28 *Let $L : Y \times \mathbb{R} \to [0, \infty]$ be a supervised loss function. Then for all $\eta \in [0, 1]$ and $\varepsilon \in (0, \infty]$ we have*

$$\delta_{\max,\bar{L}_{\text{DLD}},L}(\varepsilon, \eta) = \begin{cases} \infty & \text{if } \varepsilon > 1 - \eta \\ \inf_{t\in\mathbb{R}:(2\eta-1)\operatorname{sign} t<0} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} & \text{if } \varepsilon \le 1 - \eta. \end{cases}$$

Proof: A simple calculation shows $\mathcal{C}^*_{\bar{L}_{\text{DLD}},\eta} = 0$, and consequently we obtain $\mathcal{M}_{\bar{L}_{\text{DLD}},\eta}(\varepsilon) = \mathbb{R}$ if $\varepsilon > 1 - \eta$, and $\mathcal{M}_{\bar{L}_{\text{DLD}},\eta}(\varepsilon) = \{t \in \mathbb{R} : (2\eta - 1)\operatorname{sign} t \ge 0\}$ otherwise. ∎

With the help of the above lemma we now obtain the first main result which compares classification calibration with $\bar{L}_{\text{DLD}}$-calibration:

Theorem 4.29 *Let $L : Y \times \mathbb{R} \to [0, \infty]$ be a supervised loss function, $\eta \in [0, 1]$, and $0 \le \varepsilon \le \min\{1 - \eta, |2\eta - 1|\}$. Then we have*

$$\delta_{\max,\bar{L}_{\text{DLD}},L}(\varepsilon, \eta) \ge \delta_{\max,L_{\text{class}},L}(\varepsilon, \eta) ,$$

and consequently, L is $\bar{L}_{\text{DLD}}$-calibrated if L is classification calibrated. Moreover, if L is continuous then the above inequality becomes an equality and L is classification calibrated if and only if L is $\bar{L}_{\text{DLD}}$-calibrated.

Proof: Combining Lemma 4.1 with Lemma 4.28 yields

$$\delta_{\max,\bar{L}_{\text{DLD}},L}(\varepsilon, \eta) = \inf_{\substack{t\in\mathbb{R}: \\ (2\eta-1)\operatorname{sign} t<0}} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} \;\ge\; \inf_{\substack{t\in\mathbb{R}: \\ (2\eta-1)\operatorname{sign} t\le 0}} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} = \delta_{\max,L_{\text{class}},L}(\varepsilon, \eta) .$$

Moreover, for continuous L the assertion can be found using the continuity of $t \mapsto \mathcal{C}_{L,\eta}(t)$. ∎

By the results on classification calibrated, margin-based loss functions in [1] we immediately obtain a variety of $\bar{L}_{\mathrm{DLD}}$-calibrated losses. Furthermore, the P-instances of $\bar{L}_{\mathrm{DLD}}$ are bounded loss functions and therefore Theorem 2.8 yields

$$\mathcal{R}_{L,P}(f_n) \to \mathcal{R}^*_{L,P} \qquad \Longrightarrow \qquad \mathcal{R}_{L_{\mathrm{DLD}},\mu}(f_n) \to 0$$

whenever $P = g\mu \ominus_\rho \mu$ and L is classification calibrated. In addition, for $L := L_{\mathrm{class}}$ the proof of [28, Thm. 4] shows that the converse implication is also true.

Our next goal is to identify the *uniformly* $\bar{L}_{\mathrm{DLD}}$-calibrated losses. The following theorem gives a complete, though rather disappointing solution:

Theorem 4.30 (No uniform DLD calibration) *There exists no supervised loss function $L : Y \times \mathbb{R} \to [0,\infty]$ that is uniformly $\bar{L}_{\mathrm{DLD}}$-calibrated with respect to both $\{Q \in \mathcal{Q}_Y : Q(\{1\}) \in [0,1/2)\}$ and $\{Q \in \mathcal{Q}_Y : Q(\{1\}) \in (1/2,3/4]\}$.*

Proof: Let $L : Y \times \mathbb{R} \to [0,\infty]$ be a supervised loss function. For $\eta \in [0,1]$ we define

$$\begin{aligned} h^+(\eta) &= \inf_{t<0} \mathcal{C}_{L,\eta}(t) \\ h^-(\eta) &= \inf_{t\geq 0} \mathcal{C}_{L,\eta}(t) \end{aligned}$$

Then the functions $h^+ : [0,1] \to [0,\infty)$ and $h^- : [0,1] \to [0,\infty)$ can be defined by suprema taken over affine linear functions in $\eta \in \mathbb{R}$, and since h^+ and h^- are also finite for $\eta \in [0,1]$, we see that h^+ and h^- are continuous at every $\eta \in [0,1]$. Moreover, we have $\mathcal{C}^*_{L,\eta} = \min\{h^+(\eta), h^-(\eta)\}$ for all $\eta \in [0,1]$, and hence $\mathcal{C}^*_{L,\eta}$ is continuous in η. Let us first consider the case $\mathcal{C}^*_{L,1/2,x} = h^+(1/2)$. To this end we first observe that there exists a sequence $(t_n) \subset (-\infty,0)$ with

$$h^+\big(1/2+1/n\big) \;\leq\; \mathcal{C}_{L,1/2+1/n,x}(t_n) \;\leq\; h^+\big(1/2+1/n\big) + 1/n \tag{56}$$

for all $n \geq 1$. Moreover, our assumption $\mathcal{C}^*_{L,1/2,x} = h^+(1/2)$ yields

$$\begin{aligned} &\big|\mathcal{C}_{L,1/2+1/n,x}(t_n) - \mathcal{C}^*_{L,1/2+1/n,x}\big| \\ \leq\;& \big|\mathcal{C}_{L,1/2+1/n,x}(t_n) - h^+(1/2+1/n)\big| + \big|h^+\big(1/2+1/n\big) - h^+(1/2)\big| + \big|\mathcal{C}^*_{L,1/2,x} - \mathcal{C}^*_{L,1/2+1/n,x}\big| \end{aligned}$$

for all $n \geq 1$. By (56) and the continuity of h^+ and $\eta \mapsto \mathcal{C}^*_{L,\eta}$ we hence find

$$\lim_{n\to\infty} \big|\mathcal{C}_{L,1/2+1/n,x}(t_n) - \mathcal{C}^*_{L,1/2+1/n,x}\big| = 0\,.$$

For $\mathcal{Q} := \big\{Q \in \mathcal{Q}_Y : Q(\{1\}) \in (1/2,3/4]\big\}$ Lemma 4.28, the definition h^+, and (56) then yield

$$\delta_{\max,\bar{L}_{\mathrm{DLD}},L}(\varepsilon,\mathcal{Q}) = \inf_{\eta\in(\frac{1}{2},\frac{3}{4}]} h^+(\eta) - \mathcal{C}^*_{L,\eta} \leq \inf_{n\geq 1} \mathcal{C}_{L,1/2+1/n,x}(t_n) - \mathcal{C}^*_{L,1/2+1/n,x} = 0\,.$$

Consequently, L is not uniformly $\bar{L}_{\mathrm{DLD}}$-calibrated with respect to $\mathcal{Q}$. Finally, in the case $\mathcal{C}^*_{L,1/2,x} = h^-(1/2)$ we can analogously show that L is not uniformly $\bar{L}_{\mathrm{DLD}}$-calibrated with respect to $\{Q \in \mathcal{Q}_Y : Q(\{1\}) \in [0,1/2)\}$. ∎

Obviously the above theorem shows that there exists no uniformly $\bar{L}_{\mathrm{DLD}}$-calibrated, supervised loss function. Now recall that Theorem 2.17 showed that uniform calibration is *necessary* to establish inequalities between excess risks if essentially no assumptions on the data-generating distribution are imposed.[5] Together with Theorem 4.30 we consequently see that it is *impossible* to find

[5]Formally the result only holds for loss functions, *not template losses* However, it is quite straightforward to see that the proof of Theorem 2.17 can be easily modified to establish an analogous result for instances of template losses.

a supervised loss $L : Y \times \mathbb{R} \to [0,\infty]$ and an increasing function $\delta : [0,\infty] \to [0,\infty]$ with $\delta(0) = 0$ and $\delta(\varepsilon) > 0$, $\varepsilon > 0$, such that

$$\delta\big(\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f)\big) \;\leq\; \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} \tag{57}$$

holds for *all* μ, g, ρ, f, and $P := g\mu \ominus_\rho \mu$. However, in the DLD learning scenario we actually know μ and ρ and hence the question remains whether for certain *fixed* μ and ρ there exists a non-trivial function δ satisfying (57). In order to answer this question (negatively) we need the following elementary lemma:

Lemma 4.31 *Let $\eta \in [0,1)$ and $\rho > 0$. Furthermore, let X be a measurable space and μ be a distribution on X such that there exists measurable $A \subset X$ with $0 < \mu(A) < \min\{1, \frac{1-\eta}{\eta\rho}\}$. Then $g : X \to [0,\infty)$ defined by*

$$g(x) := \frac{\eta\rho}{1-\eta}\mathbf{1}_A(x) + \frac{1-\eta-\eta\rho\mu(A)}{(1-\eta)(1-\mu(A))}\mathbf{1}_{X\setminus A}(x) \tag{58}$$

is a density, i.e. $\|g\|_{\mathcal{L}_1(\mu)} = 1$.

Proof: Since $\mu(A) \leq \frac{1-\eta}{\eta\rho}$ we have $1-\eta-\eta\rho\mu(A) \geq 0$ and hence we actually have $g(x) \geq 0$ for all $x \in X$. Moreover, an easy calculation shows

$$\int_X g\,d\mu \;=\; \frac{\eta\rho}{1-\eta}\mu(A) + \frac{1-\eta-\eta\rho\mu(A)}{(1-\eta)(1-\mu(A))}\mu(X\setminus A) \;=\; 1\,.$$

■

With the help of the above lemma we can now show that there exists no non-trivial function δ satisfying (57) even if we fix μ and ρ:

Theorem 4.32 (No general DLD calibration inequality) *Let X be a measurable space and μ be a distribution on X such that for all $r \in [0,1]$ there exists a measurable $A \subset X$ with $\mu(A) = r$. Furthermore let $\rho > 0$, $Y := \{-1,1\}$, and $L : Y \times \mathbb{R} \to [0,\infty)$ be a supervised loss function. Then there exists no increasing function $\delta : [0,\infty] \to [0,\infty]$ with $\delta(0) = 0$ and $\delta(\varepsilon) > 0$ for $\varepsilon > 0$ such that for all measurable $g : X \to [0,\infty)$ with $\|g\|_{\mathcal{L}_1(\mu)} = 1$ and all measurable $f : X \to \mathbb{R}$ we have*

$$\delta\big(\mathcal{R}_{L_{\mathrm{DLD}},\mu}(f)\big) \;\leq\; \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}\,,$$

where $P := g\mu \ominus_\rho \mu$ and $L_{\mathrm{DLD}}(x,t) := \mathbf{1}_{(-\infty,0)}\big((g(x)-\rho)\operatorname{sign} t\big)$ for $x \in X$ and $t \in \mathbb{R}$.

Proof: Theorem 4.30 shows that L is not uniformly $\bar{L}_{\mathrm{DLD}}$-calibrated with respect to both $\{Q \in \mathcal{Q}_Y : Q(\{1\}) \in [0,1/2)\}$ and $\{Q \in \mathcal{Q}_Y : Q(\{1\}) \in (1/2,3/4]\}$. For brevity's sake we only consider the case where L is not uniformly $\bar{L}_{\mathrm{DLD}}$-calibrated with respect to $\{Q \in \mathcal{Q}_Y : Q(\{1\}) \in [0,1/2)\}$. Let us now assume that there exists a function δ in the sense of the theorem. We define $\tilde{\delta}(\varepsilon) := \delta(\frac{1+\rho}{\rho}\varepsilon)$, $\varepsilon \geq 0$. By (55) we then have

$$\tilde{\delta}\big(\mathcal{R}_{\bar{L}_{\mathrm{DLD}},P,P}(f)\big) \;\leq\; \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} \tag{59}$$

for all measurable $f : X \to \mathbb{R}$ and all P in the above sense. Let us fix an $\eta \in [0,1)$. By Lemma 4.31 there then exists a density $g : X \to [0,\infty)$ with

$$\mu\Big(\Big\{x \in X : \eta = \frac{g(x)}{g(x)+\rho}\Big\}\Big) > 0\,. \tag{60}$$

We define $P := g\mu \ominus_\rho \mu$. Then (59) together with Theorem 3.3 and $P_X = \frac{g+\rho}{1+\rho}\mu$ shows

$$\mu\Big(\big\{x \in X : \delta_{\max,\bar{L}_{\mathrm{DLD},P},L}(\varepsilon, P(.|x), x) = 0\big\}\Big) = 0\,, \qquad \varepsilon > 0.$$

Now recall that the calibration function for a P-instance of a template loss can be calculated by (28). In our case this formula yields

$$\delta_{\max,\bar{L}_{\mathrm{DLD},P},L}(\varepsilon, P(.|x), x) = \delta_{\max,\bar{L}_{\mathrm{DLD},L}}(\varepsilon, P(.|x))\,, \qquad \varepsilon > 0,\, x \in X.$$

Identifying $P(.|x)$ with $P(y = 1|x) = \frac{g(x)}{g(x)+\rho}$ and using (60) we then obtain $\delta_{\max,\bar{L}_{\mathrm{DLD}},L}(\varepsilon, \eta) > 0$ for all $\varepsilon > 0$. Since Lemma 4.28 shows that the latter also holds for $\eta = 1$ we have shown that L is $\bar{L}_{\mathrm{DLD}}$-calibrated.
Let us now fix a measurable subset $A \subset X$ with $\mu(A) = \frac{1}{2}\min\{1, \frac{1}{\rho}\}$. Furthermore, for a fixed $\eta \in [0, 1/2)$ we have $\mu(A) < \frac{1}{\rho} \leq \frac{1-\eta}{\eta\rho}$ and hence g defined by (58) is a density. Again we write $P := g\mu \ominus_\rho \mu$, and in addition we fix an $\varepsilon > 0$ with $\varepsilon \leq 1 - \eta$. Since L is DLD calibrated there then exists a $\hat{\delta} \in (0, \varepsilon)$ such that

$$\mathcal{C}_{L,\hat{\eta},x}(s) < \mathcal{C}^*_{L,\hat{\eta},x} + \hat{\delta} \qquad \Longrightarrow \qquad \mathcal{C}_{\bar{L}_{\mathrm{DLD}},\hat{\eta},x}(s) < \varepsilon\,,$$

where $\hat{\eta} := P(y = 1|x) = \frac{g(x)}{g(x)+\rho}$ for some $x \in X \setminus A$. In addition, $L(y, t) < \infty$ implies $\mathcal{C}^*_{L,\hat{\eta},x} < \infty$ and hence there *exists* an $s \in \mathbb{R}$ with $\mathcal{C}_{L,\hat{\eta},x}(s) < \mathcal{C}^*_{L,\hat{\eta},x} + \hat{\delta}$. Moreover, since $\varepsilon \leq 1 - \eta$ the definition of $\bar{L}_{\mathrm{DLD}}$ shows that this s actually satisfies $\mathcal{C}_{\bar{L}_{\mathrm{DLD}},\hat{\eta},x}(s) = 0$. For arbitrary $t \in \mathbb{R}$ we now define

$$f := t\mathbf{1}_A + s\mathbf{1}_{X\setminus A}\,.$$

Since $P(y = 1|x) = \frac{g(x)}{g(x)+\rho} = \eta$ for $x \in A$ our construction then yields

$$\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} \leq \mu(A)\big(\mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta}\big) + \big(1 - \mu(A)\big)\hat{\delta} \leq \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} + \varepsilon$$

and

$$\mathcal{R}_{\bar{L}_{\mathrm{DLD},P},P}(f) = \mu(A)\,\mathcal{C}_{\bar{L}_{\mathrm{DLD}},\eta}(t) = c_\rho\,(\mathcal{C}_{\bar{L}_{\mathrm{DLD}},\eta}(t) - \mathcal{C}^*_{\bar{L}_{\mathrm{DLD}},\eta})\,,$$

where $c_\rho := \frac{1}{2}\min\{1, \frac{1}{\rho}\}$. Combining these estimates with (59) we find

$$\tilde{\delta}\big(c_\rho\,(\mathcal{C}_{\bar{L}_{\mathrm{DLD}},\eta}(t) - \mathcal{C}^*_{\bar{L}_{\mathrm{DLD}},\eta})\big) \;\leq\; \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} + \varepsilon$$

for all $\eta \in [0, 1/2)$, $\varepsilon > 0$, and $t \in \mathbb{R}$. From the latter and Lemma 2.9 it is easy to conclude that L uniformly $\bar{L}_{\mathrm{DLD}}$-calibrated with respect to $\{Q \in \mathcal{Q}_Y : Q(\{1\}) \in [0, 1/2)\}$. However, this contradicts our initial assumption. ∎

Recall that the key idea of the proof of Theorem 4.30 was to choose η arbitrarily close to $1/2$. By finding densities g that are close to the critical level ρ on a sufficiently large set this idea was then used to prove Theorem 4.32. On the other hand if we only consider densities g that are bounded away from the level ρ then the corresponding $\eta(x) = \frac{g(x)}{g(x)+\rho}$, $x \in X$, is bounded away from the critical level $1/2$, and thus the arguments of Theorem 4.30 and Theorem 4.32 do not work. This observation has been implicitly used for the inequalities established in [28, Thm. 10]. Moreover, using Theorem 3.9 we can actually improve these inequalities slightly. Due to space constraints we only mention the result and omit the proof:

Theorem 4.33 (DLD calibration inequalities for certain densities) *Let μ be a distribution on X, $g : X \to [0,\infty]$ be a measurable function with $\|g\|_{\mathcal{L}_1(\mu)} = 1$, and $\rho > 0$. Then for $P := g\mu \ominus_\rho \mu$ the following statements hold:*

i) If there exist constants $c > 0$ and $\beta \in (0,\infty]$ with

$$\mu\big(\{x \in X : 0 < |g - \rho| < s\}\big) \leq (c\,s)^\beta \tag{61}$$

for all $s > 0$ then for all measurable $f : X \to \mathbb{R}$ we have

$$\mathcal{R}_{L_{\text{DLD}},\mu}(f) \;\leq\; 2\big((1+\rho)c\big)^{\frac{\beta}{1+\beta}}\big(\mathcal{R}_{L_{\text{class}},P}(f) - \mathcal{R}^*_{L_{\text{class}},P}\big)^{\frac{\beta}{1+\beta}}\,.$$

ii) If there exist constants $c > 0$ and $p \in (1,\infty]$ with

$$\mu\big(\{x \in X : |g - \rho| \geq s^{-1}\}\big) \leq (c\,s)^p \tag{62}$$

for all $s > 0$ then for all measurable $f : X \to \mathbb{R}$ we have

$$\mathcal{R}_{L_{\text{class}},P}(f) - \mathcal{R}^*_{L_{\text{class}},P} \;\leq\; 2\Big(\frac{p}{p-1}\Big)^{\frac{1}{p}}\frac{c}{1+\rho}\big(\mathcal{R}_{L_{\text{DLD}},\mu}(f)\big)^{\frac{p-1}{p}}.$$

Note that Condition (62) is equivalent to saying $|g - \rho| \in \mathcal{L}_{p,\infty}(\mu)$ with $\|g - \rho\|_{p,\infty} \leq c$. Consequently, if we actually have $g \in \mathcal{L}_p(\mu)$ then (62) is satisfied for $c = \|g\|_p + \rho$. Moreover, the latter condition is almost sharp since $|g - \rho| \in \mathcal{L}_{p,\infty}(\mu)$ conversely implies $g \in \mathcal{L}_{p-\varepsilon}(\mu)$ for all sufficiently small $\varepsilon > 0$.

4.5 Density Estimation

In this last subsection we apply the developed theory to the density estimation problem. To this end let us first recall that in this *unsupervised* learning problem we have a data-generating distribution on X which is of the form $g\mu$, where μ is a known distribution and g is an unknown density. The learning goal is then to estimate g. Let us assume that we have an estimate f of this density. Then the quality of this estimate is usually measured by

$$\mathcal{R}_{L_g,\mu}(f) := \int_X \big|f(x) - g(x)\big| d\mu(x) = \|f - g\|_{\mathcal{L}_1(\mu)}\,,$$

or if $g \in \mathcal{L}_p(\mu)$ is known, by $\mathcal{R}_{\bar{L}_g,p,\mu}(f) := \|f - g\|_{\mathcal{L}_p(\mu)}$. Obviously, the above performance measure $\mathcal{R}_{L_g,\mu}(.)$ is a risk with respect to the unsupervised loss function $(x,t) \mapsto |t - g(x)|$. However, like for the density level detection problem, this loss function is not accessible to us since we do not know g. Consequently, if we want to assess the quality of an estimate, we need a surrogate risk, i.e. a surrogate loss function. In order to find such a surrogate recall that a well-known heuristic for the density estimation problem is based on using additional samples drawn from μ (see e.g. [13, Chap. 14.2.4]). Let us now briefly describe and analyze this approach. To this we define $P := g\mu \ominus_1 \mu$ on $X \times Y$, $Y := \{-1, 1\}$, i.e. P is the joint distribution of the original, positively labeled samples drawn from $g\mu$, and the artificial, negatively labeled samples drawn from μ. As usual, we identify the conditional probability $P(.|x)$ with $\eta(x)$, via the relation $P(y = 1|x) = \eta(x)$, so that we have

$$\eta(x) = \frac{g(x)}{1 + g(x)} \qquad \text{and} \qquad g(x) = \frac{\eta(x)}{1 - \eta(x)}\,.$$

Moreover, we need the following definition:

Definition 4.34 *Let $L : Y \times \overline{\mathbb{R}} \to [0,\infty)$ be a supervised loss function such that for all $\eta \in [0,1)$ the set $\mathcal{M}_{L,\eta}(0^+)$ of exact minimizers contains a single element, denoted by $m(\eta)$. Then L is called a* density estimation loss function *if the map $m : [0,1) \to M := \{m(\eta) : \eta \in [0,1)\}$ defined by $\eta \mapsto m(\eta)$ is a bijection and M is a Polish space*[6].

Let us assume that we have a density estimation loss function L with associated map $m : [0,1) \to M$. Given a $t \in M$ one can then interpret $m^{-1}(t)$ as an estimate of η, and therefore the template loss

$$\tilde{L}(\eta, t) := 2(1-\eta)\Big|\frac{m^{-1}(t)}{1-m^{-1}(t)} - \frac{\eta}{1-\eta}\Big|$$

measures the quality of the estimate $\frac{m^{-1}(t)}{1-m^{-1}(t)}$ for g. Moreover, for measurable $f : X \to M$ an easy calculation shows

$$\mathcal{R}_{\tilde{L},P}(f) = \int_X 2\big(1-\eta(x)\big)\Big|\frac{m^{-1}(f(x))}{1-m^{-1}(f(x))} - \frac{\eta(x)}{1-\eta(x)}\Big|\, dP_X(x) = \int_X \Big|\frac{m^{-1}(f(x))}{1-m^{-1}(f(x))} - g(x)\Big|\, d\mu(x),$$

i.e. we have $\mathcal{R}_{\tilde{L},P}(f) = \mathcal{R}_{L_g,\mu}(\frac{m^{-1}(f)}{1-m^{-1}(f)})$. Consequently, it suffices to investigates surrogates for the template loss $\tilde{L}$. Let us begin with a negative result:

Theorem 4.35 *There exists no uniformly $\tilde{L}$-calibrated density estimation loss function L.*

Proof: Let us fix $\eta \in [0,1)$ and $\varepsilon > 0$. Then we have $\frac{2\eta+\varepsilon}{2+\varepsilon} \in [0,1)$ and consequently $t_\eta := m(\frac{2\eta+\varepsilon}{2+\varepsilon})$ is well-defined. In addition, an easy calculation shows $\tilde{L}(\eta, t_\eta) = \varepsilon$ and hence we obtain

$$\delta_{\max,\tilde{L},L}(\varepsilon,\eta) = \inf_{t\in M:\tilde{L}(\eta,t)\geq\varepsilon} \mathcal{C}_{L,\eta}(t) - \mathcal{C}^*_{L,\eta} \leq \mathcal{C}_{L,\eta}(t_\eta) - \mathcal{C}^*_{L,\eta}\,.$$

Moreover, we have

$$\begin{aligned}
\mathcal{C}_{L,\eta}(t_\eta) &= \eta L(1,t_\eta) + (1-\eta)L(-1,t_\eta) \\
&= \frac{2\eta+\varepsilon}{2+\varepsilon}L(1,t_\eta) + \Big(1-\frac{2\eta+\varepsilon}{2+\varepsilon}\Big)L(-1,t_\eta) + \frac{\varepsilon(\eta-1)}{2+\varepsilon}L(1,t_\eta) + \frac{\varepsilon(1-\eta)}{2+\varepsilon}L(-1,t_\eta) \\
&= \mathcal{C}^*_{L,\frac{2\eta+\varepsilon}{2+\varepsilon}} + \frac{\varepsilon}{2+\varepsilon}\,\mathcal{C}_{L,\eta}(t_\eta) - \frac{\varepsilon}{2+\varepsilon}\,\mathcal{C}_{L,1}(t_\eta)\,,
\end{aligned}$$

and consequently we obtain

$$\mathcal{C}_{L,\eta}(t_\eta) = \frac{2+\varepsilon}{2}\,\mathcal{C}^*_{L,\frac{2\eta+\varepsilon}{2+\varepsilon}} - \frac{\varepsilon}{2}\,\mathcal{C}_{L,1}(t_\eta) \;\leq\; \frac{2+\varepsilon}{2}\,\mathcal{C}^*_{L,\frac{2\eta+\varepsilon}{2+\varepsilon}} - \frac{\varepsilon}{2}\,\mathcal{C}^*_{L,1}\,.$$

Together with our estimate on the calibration function this yields

$$\delta_{\max,\tilde{L},L}(\varepsilon,\eta) \leq \frac{2+\varepsilon}{2}\,\mathcal{C}^*_{L,\frac{2\eta+\varepsilon}{2+\varepsilon}} - \frac{\varepsilon}{2}\,\mathcal{C}^*_{L,1} - \mathcal{C}^*_{L,\eta}\,.$$

Now recall that we have already seen in the proof of Theorem 4.30, that $\eta \mapsto \mathcal{C}^*_{L,\eta}$ is continuous on $[0,1]$ and hence we find $\lim_{\eta\to1}\delta_{\max,\tilde{L},L}(\varepsilon,\eta) = 0$. From this we easily infer the assertion. ∎

[6]This is only a technical assumption which is satisfied for all commonly used surrogates.

Remark 4.36 At first glance one is tempted to conclude from Theorem 4.35 and Theorem 2.17 that there cannot exist a general inequality between the risk $\mathcal{R}_{L_g,\mu}(\frac{m^{-1}(f)}{1-m^{-1}(f)})$ and an excess surrogate risk $\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}$. However, in the case of density estimation we are actually not interested in all distributions P on $X \times Y$, but only in distributions of the form $g\mu \ominus_1 \mu$ with g being a *density* with respect to μ, and therefore Theorem 2.17 does not apply *directly*. Nevertheless, Theorem 4.35 can be used to show that no density estimation loss function L allow a general inequality between $\mathcal{R}_{L_g,\mu}(\frac{m^{-1}(f)}{1-m^{-1}(f)})$ and its excess risk $\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}$. To see let $X := \{1,2\}$, μ be a distribution on X and $g : X \to [0,\infty)$ be a density with $g(2) = 0$. Note that μ is uniquely determined by $\mu_1 := \mu(\{1\})$, and since g is a density with respect to μ we have $g_1 := g(1) = 1/\mu_1$. Using our standard notations this yields $\eta_1 := \eta(1) \geq 1/2$ and $\eta(2) = 0$. Now note that for $f : X \to \mathbb{R}$ with $m^{-1}(f(2)) = 0 = g(2)$ we have

$$\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} = \frac{\mathcal{C}_{L,\eta_1}(f_1) - \mathcal{C}^*_{L,\eta_1}}{2\eta_1} \leq \mathcal{C}_{L,\eta_1}(f_1) - \mathcal{C}^*_{L,\eta_1}\,,$$

where $f_1 := f(1)$. Moreover, for such f we also have

$$\mathcal{R}_{L_g,\mu}\Big(\frac{m^{-1}(f)}{1-m^{-1}(f)}\Big) = \frac{1-\eta_1}{\eta_1} \cdot \Big|\frac{m^{-1}(f_1)}{1-m^{-1}(f_1)} - \frac{\eta_1}{1-\eta_1}\Big| = \frac{\tilde{L}(\eta_1,f_1)}{2\eta_1}.$$

Now assume that we have a function $\delta : [0,\infty] \to [0,\infty]$ and an inequality in the sense of Theorem 2.17. Then this inequality in particular holds for $f : X \to \mathbb{R}$ with $m^{-1}(f(2)) = 0$, and hence we find

$$\delta\Big(\frac{\tilde{L}(\eta_1,f_1)}{2}\Big) \leq \delta\Big(\frac{\tilde{L}(\eta_1,f_1)}{2\eta_1}\Big) = \delta\bigg(\mathcal{R}_{L_g,\mu}\Big(\frac{m^{-1}(f)}{1-m^{-1}(f)}\Big)\bigg) \leq \mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P} \leq \mathcal{C}_{L,\eta_1}(f_1) - \mathcal{C}^*_{L,\eta_1}\,.$$

Since this inequality holds for all $\eta_1 \in [1/2,1)$ and all $f_1 \in \mathbb{R}$, we easily see that L is uniformly $\tilde{L}$-calibrated with respect to $\{\eta : 1/2 \leq \eta < 1\}$. However, we have seen in (the proof of) Theorem 4.35 that this uniform calibration is impossible, and consequently no general inequalities are possible in the above sense.

After these disappointing results let us finally present a positive result for convex loss functions:

Theorem 4.37 *Let $L : Y \times \overline{\mathbb{R}} \to [0,\infty)$ be density estimation loss function which is convex in its second argument. Then L is $\tilde{L}$-calibrated.*

Proof: Repeating the proof of [26, Lem. 20] we see that the map $m : [0,1) \to M$ is monotone, and since it is also invertible, it must be strictly monotone. Moreover, we obviously have $\mathcal{M}_{L,Q,x}(0^+) \subset \mathcal{M}_{\tilde{L},Q,x}(0^+)$. Using Lemma 2.11 and the ideas of its proof we then obtain the assertion. ∎

Remark 4.38 Let $L : Y \times \overline{\mathbb{R}} \to [0,\infty)$ be density estimation loss function that is $\tilde{L}$-calibrated. Repeating the proof of Theorem 3.16 we then see that

$$\frac{m^{-1}(f_n)}{1-m^{-1}(f_n)} \to g \quad \text{in probability}$$

whenever $f_n : X \to M$ is a sequence of measurable functions with $\mathcal{R}_{L,P}(f_n) \to \mathcal{R}^*_{L,P}$. In particular, *every* universally L-risk consistent method is also universally consistent in the above weak form of the density estimation problem. Moreover, if g is a *bounded* density, then it is not hard to derive consistency with respect to $\mathcal{R}_{L_g,\mu}(.)$. Finally, we like to mention that additional results can be obtained by computing the calibration function $\delta_{\max,\tilde{L},L}(.,.)$ for specific L, but because of space constraints we omit the details.

5 Proofs for the Results of Section 2 and 3

Proof of Lemma 2.5: For $n \in \mathbb{N}$ let us fix an $\alpha_n \in \mathcal{A}$ that satisfies (2) for $\varepsilon := 1/n$. Then for all $x \in X$ we have

$$\mathcal{C}^*_{L,P(.|x),x} = \inf_{\alpha\in\mathcal{A}} \mathcal{C}_{L,P(.|x),x}(\alpha) = \inf_{n\in\mathbb{N}} \mathcal{C}_{L,P(.|x),x}(\alpha_n)\,,$$

and hence $x \mapsto \mathcal{C}^*_{L,P(.|x),x}$ is measurable. Furthermore for all $n \geq 1$ we have

$$\begin{aligned}\mathcal{R}^*_{L,P} \leq \int_X \mathcal{C}_{L,P(.|x),x}(\alpha_n)\, dP_X(x) \leq \int_X \mathcal{C}^*_{L,P(.|x),x}\, dP_X(x) + \frac{1}{n} &\leq \inf_{\alpha\in\mathcal{A}} \int_X \mathcal{C}_{L,P(.|x),x}(\alpha)\, dP_X(x) + \frac{1}{n} \\ &\leq \mathcal{R}^*_{L,P} + \frac{1}{n}\,,\end{aligned}$$

and hence we obtain the second assertion if we let $n \to \infty$. ∎

Proof of Theorem 2.8: For brevity's sake we write $\mathcal{C}_{1,x}(\alpha) := \mathcal{C}_{L_1,P(.|x),x}(\alpha) - \mathcal{C}^*_{L_1,P(.|x),x}$ and $\mathcal{C}_{2,x}(\alpha) := \mathcal{C}_{L_2,P(.|x),x}(\alpha) - \mathcal{C}^*_{L_2,P(.|x),x}$ for $x \in X$, $\alpha \in \mathcal{A}$. Furthermore let us fix an $\varepsilon > 0$ and write $h(x) := \delta(\varepsilon, x)$, $x \in X$. Then for all $x \in X$, and all $\alpha \in \mathcal{A}$ with $\mathcal{C}_{1,x}(\alpha) \geq \varepsilon$ we have $\mathcal{C}_{2,x}(\alpha) \geq h(x)$, and hence we obtain

$$\mathcal{R}_{L_2,P}(\alpha) - \mathcal{R}^*_{L_2,P} \;=\; \int_X \mathcal{C}_{2,x}(\alpha)\, dP_X(x) \;\geq\; \int_{\mathcal{C}_{1,x}(\alpha)\geq\varepsilon} h(x) dP_X(x)$$

for all $\alpha \in \mathcal{A}$. Furthermore, since $h(x) > 0$ holds for all $x \in X$, the measure $\nu := b\, P_X$ is absolutely continuous with respect to $\mu := h\, P_X$ and consequently there exists a $\delta > 0$ such that $\nu(A) \leq \varepsilon$ for all measurable $A \subset X$ with $\mu(A) \leq \delta$. For $\alpha \in \mathcal{A}$ with $\mathcal{R}_{L_2,P}(\alpha) - \mathcal{R}^*_{L_2,P} \leq \delta$ and $A := \{x \in X : \mathcal{C}_{1,x}(\alpha) \geq \varepsilon\}$ the above considerations thus show

$$\begin{aligned}\mathcal{R}_{L_1,P}(\alpha) - \mathcal{R}^*_{L_1,P} = \int_{\mathcal{C}_{1,x}(\alpha)\geq\varepsilon} \mathcal{C}_{1,x}(\alpha) dP_X(x) + \int_{\mathcal{C}_{1,x}(\alpha)<\varepsilon} \mathcal{C}_{1,x}(\alpha) dP_X(x) &\leq \int_A b(x) dP_X(x) + \varepsilon \\ &\leq 2\varepsilon\,.\end{aligned}$$

From this we easily get the assertion. ∎

Proof of Lemma 2.9: Let us first assume $\mathcal{C}^*_{L_2,Q,x} = \infty$. Then we have $\delta_{\max}(\varepsilon, Q, x) = \infty$ and hence $ii)$ is trivially satisfied. Furthermore we have $\mathcal{M}_{L_2,Q,x}(\delta_{\max}(\varepsilon, Q, x)) = \emptyset$, and hence we obtain $i)$. Let us now assume $\mathcal{C}^*_{L_2,Q,x} < \infty$. Then for $\alpha \in \mathcal{M}_{L_2,Q,x}(\delta_{\max}(\varepsilon, Q, x))$ we have

$$\mathcal{C}_{L_2,Q,x}(\alpha) - \mathcal{C}^*_{L_2,Q,x} \;<\; \delta_{\max}(\varepsilon, Q, x) \;=\; \inf_{\substack{\alpha'\in\mathcal{A} \\ \alpha'\notin\mathcal{M}_{L_1,Q,x}(\varepsilon)}} \mathcal{C}_{L_2,Q,x}(\alpha') - \mathcal{C}^*_{L_2,Q,x},$$

which shows $\alpha \in \mathcal{M}_{L_1,Q,x}(\varepsilon)$. For the proof of the second assertion let us fix a δ with $\delta > \delta_{\max}(\varepsilon, Q, x)$. By definition this means

$$\inf_{\substack{\alpha\in\mathcal{A} \\ \alpha\notin\mathcal{M}_{L_1,Q,x}(\varepsilon)}} \mathcal{C}_{L_2,Q,x}(\alpha) - \mathcal{C}^*_{L_2,Q,x} \;=\; \delta_{\max}(\varepsilon, Q, x) \;<\; \delta,$$

and hence there exists a $\alpha \in \mathcal{M}_{L_2,Q,x}(\delta)$ with $\alpha \notin \mathcal{M}_{L_1,Q,x}(\varepsilon)$. Finally, in order to establish (9) let us fix a $\alpha \in \mathcal{A}$. Then for $\varepsilon := \mathcal{C}_{L_1,Q,x}(\alpha) - \mathcal{C}^*_{L_1,Q,x}$ we have $\alpha \notin \mathcal{M}_{L_1,Q,x}(\varepsilon)$ which implies $\alpha \notin \mathcal{M}_{L_2,Q,x}(\delta_{\max}(\varepsilon, Q, x))$ by $i)$. Since the latter means

$$\mathcal{C}_{L_2,Q,x}(\alpha) \geq \mathcal{C}^*_{L_2,Q,x} + \delta_{\max}(\varepsilon, Q, x) = \mathcal{C}^*_{L_2,Q,x} + \delta_{\max}\left(\mathcal{C}_{L_1,Q,x}(\alpha) - \mathcal{C}^*_{L_1,Q,x}, Q, x\right)$$

we obtain the assertion. ∎

Proof of Lemma 2.11: Since $\mathcal{M}_{L_1,Q,x}(\varepsilon)$ is an interval and $\mathcal{C}_{L_2,Q,x}(.) : \mathbb{R} \to [0,\infty)$ is continuous we obtain

$$\delta_{\max}(\varepsilon, Q, x) = \min\Big\{ \inf_{\alpha \le \inf \mathcal{M}_{L_1,Q,x}(\varepsilon)} \mathcal{C}_{L_2,Q,x}(\alpha), \inf_{\alpha \ge \sup \mathcal{M}_{L_1,Q,x}(\varepsilon)} \mathcal{C}_{L_2,Q,x}(\alpha) \Big\} - \mathcal{C}^*_{L_2,Q,x} \,.$$

Moreover, for $\alpha \in \mathbb{R}$ with $\alpha \notin \mathcal{M}_{L_1,Q,x}(\varepsilon)$ we have $\alpha \notin \mathcal{M}_{L_1,Q,x}(0^+)$, and consequently we find $\alpha \notin \mathcal{M}_{L_2,Q,x}(0^+)$. Now, it is easy to check that the map $\alpha \mapsto \mathcal{C}_{L_2,Q,x}(\alpha)$ is convex, and consequently it is strictly decreasing on $(-\infty, \inf \mathcal{M}_{L_2,Q,x}(0^+)]$ and strictly increasing on $[\sup \mathcal{M}_{L_2,Q,x}(0^+), \infty)$. Combining all observations we obtain the assertion. ∎

Proof of Theorem 2.13: Let us use the shorthands $\mathcal{C}_{1,x}(\alpha)$ and $\mathcal{C}_{2,x}(\alpha)$ defined in the proof of Theorem 2.8. Assumption (12) together with Lemma 2.9 and $\mathcal{R}^*_{L_1,P} < \infty$, $\mathcal{R}^*_{L_2,P} < \infty$ then gives

$$\delta\big(\mathcal{C}_{1,x}(\alpha)\big) \;\le\; \mathcal{C}_{2,x}(\alpha) \tag{63}$$

for all $x \in \tilde{X}$ and all $\alpha \in \mathcal{A}$. For $\alpha \in \mathcal{A}$ with $\mathcal{R}_{L_1,P}(\alpha) < \infty$ Jensen's inequality together with the definition of B_α, $\delta^{**}_{B_\alpha}(.) \le \delta(.)$, and (63), now yields

$$\delta^{**}_{B_\alpha}\big(\mathcal{R}_{L_1,P}(\alpha) - \mathcal{R}^*_{L_1,P}\big) \le \int_X \delta^{**}_{B_\alpha}\big(\mathcal{C}_{1,x}(\alpha)\big) dP_X(x) \le \int_X \mathcal{C}_{2,x}(\alpha)\, dP_X(x) = \mathcal{R}_{L_2,P}(\alpha) - \mathcal{R}^*_{L_2,P} \,.$$

Finally, for $\alpha \in \mathcal{A}$ with $\mathcal{R}_{L_1,P}(\alpha) = \infty$ we have $B_\alpha = \infty$. If $\delta^{**}_\infty(\infty) = 0$ there is nothing to prove, and hence let us assume $\delta^{**}_\infty(\infty) > 0$. Then observe that because of $0 = \delta(0) = \delta^{**}_\infty(0)$ and its convexity the function δ^{**} is increasing. Consequently, if there is a $t_0 > 0$ with $\delta^{**}_\infty(t_0) = \infty$ we obviously have $t \le c_1 \delta^{**}_\infty(t) + c_2$ for $c_1 := 1$, $c_2 := t_0$ and all $t \in [0,\infty]$. On the other hand, if δ^{**} is finite on $[0,\infty)$ then there exists a $t_0 \ge 0$ and a $c_0 > 0$ such that the (Lebesgue)-almost surely defined derivative of δ^{**}_∞ satisfies $(\delta^{**}_\infty)'(t) \ge c_0$ for almost all $t \ge t_0$. By Lebesgue's version of the fundamental theorem of calculus (see e.g. the Theorems 26-28 in Chapter X of [11] or the Theorems 271, 269, and 274 in [16]) we then also find constants $c_1, c_2 \in (0,\infty)$ with $t \le c_1 \delta^{**}_\infty(t) + c_2$ for all $t \in [0,\infty]$. In both cases (63) consequently yields

$$\infty = \int_X \mathcal{C}_{1,x}(\alpha) dP_X(x) \le c_1 \int_X \delta^{**}_\infty\big(\mathcal{C}_{1,x}(\alpha)\big) dP_X(x) + c_2 \le c_1\big(\mathcal{R}_{L_2,P}(\alpha) - \mathcal{R}^*_{L_2,P}\big) + c_2 \,,$$

and hence we have $\mathcal{R}_{L_2,P}(\alpha) - \mathcal{R}^*_{L_2,P} = \infty$. This shows the assertion. ∎

Proof of Lemma 2.16: In order to show the first assertion we fix an $\varepsilon > 0$. If $\delta_{\max}(\varepsilon, \mathcal{Q}) = 0$ there is nothing to prove and hence we assume $\delta_{\max}(\varepsilon, \mathcal{Q}) > 0$ without loss of generality. Then there exists a strictly positive sequence (δ_n) with $\delta_n \nearrow \delta_{\max}(\varepsilon, \mathcal{Q})$ for $n \to \infty$, and $\mathcal{M}_{L_2,Q,x}(\delta_n) \subset \mathcal{M}_{L_1,Q,x}(\varepsilon)$ for all $n \in \mathbb{N}$. Now let us fix an $\alpha \in \mathcal{M}_{L_2,Q,x}(\delta_{\max}(\varepsilon, \mathcal{Q}))$. Then we have $\mathcal{C}_{L_2,Q,x}(\alpha) - \mathcal{C}^*_{L_2,Q,x} < \delta_{\max}(\varepsilon, \mathcal{Q})$ and hence there exists an $n \in \mathbb{N}$ with $\mathcal{C}_{L_2,Q,x}(\alpha) - \mathcal{C}^*_{L_2,Q,x} < \delta_n$. Obviously this implies $\alpha \in \mathcal{M}_{L_2,Q,x}(\delta_n) \subset \mathcal{M}_{L_1,Q,x}(\varepsilon)$, i.e. we have shown the first assertion.
In order to prove the second assertion we write $\delta(\varepsilon) := \inf_{x \in X} \delta_{\max}(\varepsilon, Q, x)$ for $\varepsilon > 0$. Since $\delta(\varepsilon) \le \delta_{\max}(\varepsilon, Q, x)$ for all $x \in X$, $Q \in \mathcal{Q}$, and $\varepsilon > 0$, we then obtain

$$\mathcal{M}_{L_2,Q,x}(\delta(\varepsilon)) \subset \mathcal{M}_{L_2,Q,x}(\delta_{\max}(\varepsilon, Q, x)) \subset \mathcal{M}_{L_1,Q,x}(\varepsilon) \,.$$

This shows $\delta(\varepsilon) \le \delta_{\max}(\varepsilon, \mathcal{Q})$. In order to prove the converse inequality let us assume $\delta(\varepsilon) < \delta_{\max}(\varepsilon, \mathcal{Q})$ for some $\varepsilon > 0$. Then there exist $Q \in \mathcal{Q}$ and $x \in X$ with $\delta_{\max}(\varepsilon, Q, x) < \delta_{\max}(\varepsilon)$. However, the already proved first assertion shows $\mathcal{M}_{L_2,Q,x}(\delta_{\max}(\varepsilon, \mathcal{Q})) \subset \mathcal{M}_{L_1,Q,x}(\varepsilon)$, and hence we find a contradiction by *ii)* in Lemma 2.9. ∎

Proof of Theorem 2.17: Let us fix an $x \in X$ and a $Q \in \mathcal{Q}$. Furthermore, let P be the distribution on $X \times Y$ with $P_X = \delta_{\{x\}}$, where $\delta_{\{x\}}$ is the Dirac measure in x, and $P(.|x) = Q$. Then P is of type $\mathcal{Q}$ and we have $\mathcal{R}_{L_i,P}(\alpha) = \mathcal{C}_{L_i,Q,x}(\alpha)$ and $\mathcal{R}^*_{L_i,P} = \mathcal{C}^*_{L_i,Q,x} < \infty$ for $i = 1, 2$ and all measurable $\alpha \in \mathcal{A}$. Consequently our assumption yields

$$\delta\big(\mathcal{C}_{L_1,Q,x}(\alpha) - \mathcal{C}^*_{L_1,Q,x}\big) \;\le\; \mathcal{C}_{L_2,Q,x}(\alpha) - \mathcal{C}^*_{L_2,Q,x}$$

for all $\alpha \in \mathcal{A}$. Now let $\varepsilon > 0$ and $\alpha \in \mathcal{M}_{L_2,Q,x}(\delta(\varepsilon))$. Then we have $\mathcal{C}_{L_2,Q,x}(\alpha) - \mathcal{C}^*_{L_2,Q,x} < \delta(\varepsilon)$ and hence the above inequality yields $\delta(\mathcal{C}_{L_1,Q,x}(\alpha) - \mathcal{C}^*_{L_1,Q,x}) < \delta(\varepsilon)$. Since δ is monotonically increasing the latter shows $\mathcal{C}_{L_1,Q,x}(\alpha) - \mathcal{C}^*_{L_1,Q,x} < \varepsilon$, i.e. we have found $\mathcal{M}_{L_2,Q,x}(\delta(\varepsilon)) \subset \mathcal{M}_{L_1,Q,x}(\varepsilon)$. Lemma 2.9 then shows $\delta(\varepsilon) \le \delta_{\max}(\varepsilon, Q, x)$, and hence L_2 is uniformly L_1-calibrated with respect to $\mathcal{Q}$. ∎

Proof of Theorem 2.18: Let us use the shorthands $\mathcal{C}_{1,x}(\alpha)$ and $\mathcal{C}_{2,x}(\alpha)$ defined in the proof of Theorem 2.8. Then our assumption on $\delta_{\max}(\varepsilon, x)$ together with (9) yields $(\mathcal{C}_{2,x}(\alpha))^q b(x) \le \mathcal{C}_{1,x}(\alpha)$ for all $x \in X$, $\alpha \in \mathcal{A}$, and hence using Hölder's inequality with r defined by $q = \frac{1}{r} + 1$ gives

$$\mathcal{R}_{L_1,P}(\alpha) - \mathcal{R}^*_{L_1,P} \le \left(\int_X \big(b(x)\big)^{-\frac{1}{q}} \big(\mathcal{C}_{1,x}(\alpha)\big)^{\frac{1}{q}} \, dP_X(x)\right)^{\frac{q}{q}} \le \left(\int_X b^{-r} dP_X\right)^{\frac{1}{qr}} \left(\int_X \mathcal{C}_{1,x}(\alpha) \, dP_X(x)\right)^{\frac{1}{q}}.$$

Since $p \ge r$ we then find the assertion. ∎

For the proof of Theorem 3.2 we need a method for finding measurable selections from multi-valued maps of specific forms. We begin with some basics: Let X, Y, and Z be measurable spaces, $h : X \times Z \to Y$ and $A \subset Y$ be measurable, and

$$\begin{array}{rcl} F : X & \to & 2^Z \\ x & \mapsto & \big\{z \in Z : h(x,z) \in A\big\}. \end{array} \tag{64}$$

Furthermore we write

$$\begin{array}{rcll} \operatorname{Dom} F & := & \big\{x \in X : F(x) \neq \emptyset\big\}, & \\ \operatorname{Gr} F & := & \big\{(x,z) \in X \times Z : z \in F(x)\big\}, & \\ F^{-1}(B) & := & \big\{x \in X : F(x) \cap B \neq \emptyset\big\} & B \subset Z. \end{array}$$

Note that we have $\operatorname{Dom} F = F^{-1}(Z)$ and $\operatorname{Gr} F = \{(x,z) \in X \times Z : h(x,z) \in A\}$, so that in particular $\operatorname{Gr} F$ is measurable. Furthermore, if $\pi_X : X \times Z \to X$ denotes the projection onto X then we have

$$\operatorname{Dom} F = \big\{x \in X : \exists z \in Z \text{ with } h(x,z) \in A\big\} = \pi_X\Big(\big\{(x,z) \in X \times Z : h(x,z) \in A\big\}\Big) = \pi_X\big(\operatorname{Gr} F\big).$$

Now the following result provides a sufficient condition under which $\operatorname{Dom} F$ is measurable and F admits measurable selections.

Lemma 5.1 *Let $(X, \mathcal{X})$ be a complete measurable space, Z be a Polish space equipped with its Borel σ-algebra, and Y be a measurable space. Furthermore let $h : X \times Z \to Y$ be a measurable map, $A \subset Y$ be measurable and $F : X \to 2^Z$ be defined by (64). Then the following are true:*

i) $\operatorname{Dom} F$ *is measurable.*

ii) There exists a sequence of measurable functions $f_n : X \to Z$ such that for all $x \in \operatorname{Dom} F$ the set $\{f_n(x) : n \in \mathbb{N}\}$ is dense in $F(x)$.

iii) Let $\varphi : X \times Z \to [0,\infty]$ be measurable and $\psi : X \to [0,\infty]$ be defined by

$$\psi(x) := \inf_{z \in F(x)} \varphi(x,z)\,, \qquad\qquad x \in X. \tag{65}$$

Then ψ is measurable. Furthermore, for all $n \geq 1$ there exists a measurable $f_n : X \to Z$ such that for all $x \in \operatorname{Dom} F$ we have $f_n(x) \in F(x)$ and $\varphi(x, f_n(x)) \leq \psi(x) + 1/n$, and consequently

$$\psi(x) = \inf_{n \in \mathbb{N}} \varphi\big(x, f_n(x)\big)\,.$$

In addition, if the infimum in (65) is attained for all $x \in \operatorname{Dom} X$ then there exists a measurable function $f : X \to Z$ with $\psi(x) = \varphi(x, f(x))$ for all $x \in \operatorname{Dom} F$.

Proof: *i*). Let us first recall that the so-called projection theorem [7, Thm. III.23, p. 75] ensures $\pi_X(B) \in \mathcal{X}$ for all $B \in \mathcal{X} \otimes \mathcal{B}(Z)$. Now the assertion directly follows from $\operatorname{Dom} F = \pi_X(\operatorname{Gr} F)$ and the projection theorem.
ii). Let $\tilde{\mathcal{X}}$ be the trace σ-algebra of $\mathcal{X}$ on $\operatorname{Dom} F$. Then it is easy to see that $\tilde{\mathcal{X}}$ is a complete σ-algebra. Furthermore, $F_{|\operatorname{Dom} F} : \operatorname{Dom} F \to 2^Z$ obviously maps $\operatorname{Dom} F$ to non-empty subsets of Z and in addition we have

$$\operatorname{Gr}(F_{|\operatorname{Dom} F}) = \big\{(x,z) \in \operatorname{Dom} F \times Z : z \in F(x)\big\} = \operatorname{Gr} F \cap \big(\operatorname{Dom} F \times Z\big)\,,$$

and hence $\operatorname{Gr}(F_{|\operatorname{Dom} F})$ is measurable. Now Aumann's selection theorem in the form of [7, Thm. III.22, p. 74] gives a sequence of $\tilde{\mathcal{X}}$-measurable functions $\tilde{f}_n : \operatorname{Dom} F \to \mathcal{T}$ such that $\{f_n(x) : n \in \mathbb{N}\}$ is dense in $F(x)$ for all $x \in \operatorname{Dom} F$. Extending these functions to measurable functions $f_n : X \to \mathcal{T}$ gives the assertion.
iii) The measurability follows from [7, Lem. III.39, p. 86]. Furthermore, on the measurable set $\{x \in X : \psi(x) = \infty\}$ there is nothing to prove and hence we may restrict our considerations to $\operatorname{Dom} F \cap \{x \in X : \psi(x) < \infty\}$ equipped with the trace σ-algebra of $\mathcal{X}$. Then the existence of f_n is shown in [7, p. 87]. Finally, the existence of a measurable $f : X \to Z$ is shown in [7, p. 86]. ■

Proof of Theorem 3.2: First note that the measurability of $(x,t) \mapsto \mathcal{C}_{L,P(.|x),x}(t)$ can be shown using standard arguments.
i). If $\mathcal{C}^*_{L,P(.|x),x} = \infty$ for some $x \in X$, then (2) cannot be true for this x, and hence $\hat{L}$ is not P-minimizable. To see the converse implication observe that the multivalued map $F : X \to 2^{\mathcal{T}}$, defined by $F(x) := \mathcal{T}$ for all $x \in X$, is obviously of the form (64) for arbitrary measurable h and $A := \mathcal{T}$. We write $\varphi(x,t) := \mathcal{C}_{L,P(.|x),x}(t)$ for all $x \in X$, $t \in \mathcal{T}$, so that $\varphi : X \times \mathcal{T} \to [0,\infty]$ is measurable. Then we have

$$\mathcal{C}^*_{L,P(.|x),x} = \inf_{t \in F(X)} \varphi(x,t) \tag{66}$$

for all $x \in X$, and consequently Lemma 5.1 shows that for all $n \geq 1$ there exists a measurable function $f_n : X \to \mathcal{T}$ with

$$\mathcal{C}_{L,P(.|x),x}(f_n(x)) \;=\; \varphi\big(x, f_n(x)\big) \leq\; \mathcal{C}^*_{L,P(.|x),x} + 1/n \;<\; \mathcal{C}^*_{L,P(.|x),x} + 2/n$$

for all $x \in \operatorname{Dom} F = X$. From this we easily conclude that $\hat{L}$ is P-minimizable.
ii). This assertion also follows from Lemma 5.1 since $\mathcal{M}_{L,P(.|x),x}(0^+) \neq \emptyset$ for all $x \in \mathcal{X}$ ensures that the infimum in (66) is attained for all $x \in X$. ■

Proof of Theorem 3.3: Without loss of generality we may assume $\mathcal{C}^*_{L_i,P(.|x),x} < \infty$ for all $x \in X$ and $i = 1, 2$. To show the measurability of $x \mapsto \delta_{\max}(\varepsilon, P(.|x), x)$ we equip $[0, \infty]$ with the Borel σ-algebra, write $A := [\varepsilon, \infty]$, and define $h : X \times \mathcal{T} \to [0, \infty]$ by

$$h(x,t) := \mathcal{C}_{L_1,P(.|x),x}(t) - \mathcal{C}^*_{L_1,P(.|x),x}\,, \qquad (x,t) \in X \times \mathcal{T}.$$

Then h is measurable and for $F : X \to 2^{\mathcal{T}}$ defined by (64) we have $\mathcal{T} \setminus \mathcal{M}_{L_1,P(.|x),x}(\varepsilon) = F(x)$ for all $x \in X$. Furthermore, $\varphi : X \times \mathcal{T} \to [0, \infty]$ defined by

$$\varphi(x,t) := \mathcal{C}_{L_2,P(.|x),x}(t) - \mathcal{C}^*_{L_2,P(.|x),x}\,, \qquad (x,t) \in X \times \mathcal{T},$$

is also measurable. For all $x \in X$ our construction yields

$$\delta_{\max}(\varepsilon, P(.|x), x) = \inf_{t \in F(x)} \varphi(x,t)\,,$$

and consequently we obtain the measurability of $x \mapsto \delta_{\max}(\varepsilon, P(.|x), x)$ by Lemma 5.1.
Let us now assume that there exists an $\varepsilon > 0$ such that $B := \{x \in X : \delta_{\max}(\varepsilon, x) = 0\}$ satisfies $P_X(B) > 0$. With the above notations we obviously have $B \subset \operatorname{Dom} F$. Moreover for all $n \geq 1$ Lemma 5.1 gives us a measurable function $f^{(1)}_n : X \to \mathcal{T}$ with

$$\mathcal{C}_{L_2,P(.|x),x}(f^{(1)}_n(x)) - \mathcal{C}^*_{L_2,P(.|x),x} \;\leq\; \frac{1}{n} \qquad \text{and} \qquad \mathcal{C}_{L_1,P(.|x),x}(f^{(1)}_n) - \mathcal{C}^*_{L_1,P(.|x),x} \;\geq\; \varepsilon$$

for all $x \in B$. Furthermore, since L_2 is P-minimizable there also exist measurable functions $f^{(2)}_n : X \to \mathcal{T}$ with

$$\mathcal{C}_{L_2,P(.|x),x}(f^{(2)}_n(x)) - \mathcal{C}^*_{L_2,P(.|x),x} \;\leq\; \frac{1}{n}$$

for all $x \in X$. We define $f_n : X \to \mathcal{T}$ by $f_n(x) := f^{(1)}_n(x)$ if $x \in B$, and $f_n(x) := f^{(2)}_n(x)$ otherwise. Then f_n is measurable and our construction yields both $\lim_{n\to\infty} \mathcal{R}_{L_2,P}(f_n) = \mathcal{R}^*_{L_2,P}$ and

$$\mathcal{R}_{L_1,P}(f) - \mathcal{R}^*_{L_1,P} \geq \int_B \mathcal{C}_{L_1,P(.|x),x}(t) - \mathcal{C}^*_{L_1,P(.|x),x}\, dP_X(x) \geq \varepsilon P_X(B)\,.$$

From this we easily obtain the assertion. ∎

Proof of Theorem 3.6: Since $\delta_{\max}(., \mathcal{Q})$ is monotonously increasing, the set

$$U := \big\{\varepsilon > 0 : \delta_{\max}(., \mathcal{Q}) \text{ not continuous in } \varepsilon\big\}$$

is at most countable, and hence there exists a sequence (ε_n) with $\{\varepsilon_n : n \in \mathbb{N}\} = U \cup \{r \in Q : r > 0\}$. Then for all $n, m \geq 1$ there exists a distribution $Q_{n,m} \in \mathcal{Q}$ with

$$\frac{1}{m} + \delta_{\max}(\varepsilon_n, \mathcal{Q}) > \delta_{\max}(\varepsilon_n, Q_{n,m}) = \inf_{\substack{t \in \mathcal{T} \\ t \notin \mathcal{M}_{L_1,Q_{n,m}}(\varepsilon_n)}} \mathcal{C}_{L_2,Q_{n,m}}(t) - \mathcal{C}^*_{L_2,Q_{n,m}}\,.$$

Therefore for all $n, m \geq 1$ there exist $t_{n,m} \in \mathcal{T}$ with $t_{n,m} \notin \mathcal{M}_{L_1,Q_{n,m}}(\varepsilon_n)$ and

$$\mathcal{C}_{L_2,Q_{n,m}}(t_{n,m}) - \mathcal{C}^*_{L_2,Q_{n,m}} < \delta_{\max}(\varepsilon_n, \mathcal{Q}) + \frac{1}{m}\,.$$

Now, let $A_{n,m} \subset X$, $n \geq 1$, be according to our assumption. Note that without loss of generality we can additionally assume $X = \bigcup_{n,m} A_{n,m}$. Consequently, for $x \in X$ there exists a unique

$(n_x, m_x) \in \mathbb{N} \times \mathbb{N}$ with $x \in A_{n_x,m_{x,m_x}}$. Furthermore, let P be the distribution on $X \times Y$ which is defined by the conditions $P_X = \mu$ and $P(.|x) = Q_{n_x,m_x}$, $x \in X$. Now let us fix an $n \geq 1$. Then for all $x \in \bigcup_m A_{n,m}$ our construction gives $t_{n,m_x} \notin \mathcal{M}_{L_1,P(.|x),x}(\varepsilon_n)$, and hence we obtain

$$\delta_{\max}(\varepsilon_n, P) \leq \mathcal{C}_{L_2,Q_{n,m_x}}(t_{n,m_x}) - \mathcal{C}^*_{L_2,Q_{n,m_x}} < \delta_{\max}(\varepsilon_n, \mathcal{Q}) + \frac{1}{m_x}\,.$$

Minimizing over $x \in \bigcup_m A_{n,m}$ then gives $\delta_{\max}(\varepsilon_n, P) \leq \delta_{\max}(\varepsilon_n, \mathcal{Q})$ for all $n \in \mathbb{N}$, and consequently we have $\delta_{\max}(\varepsilon_n, P) = \delta_{\max}(\varepsilon_n, \mathcal{Q})$ for all $n \in \mathbb{N}$. In particular, this shows $\delta_{\max}(\varepsilon, P) = \delta_{\max}(\varepsilon, \mathcal{Q})$ for all $\varepsilon \in U$. Now let $\varepsilon > 0$ with $\varepsilon \notin U$. There there exists a sub-sequence (ε_{n_k}) of (ε_n) with $\varepsilon_{n_k} \searrow \varepsilon$ for $k \to \infty$. This gives $\delta_{\max}(\varepsilon_{n_k}, P) = \delta_{\max}(\varepsilon_{n_k}, \mathcal{Q}) \to \delta_{\max}(\varepsilon, \mathcal{Q})$, and hence we have

$$\delta_{\max}(\varepsilon, \mathcal{Q}) \geq \inf_{k \geq 1} \delta_{\max}(\varepsilon_{n_k}, P) \geq \inf_{\varepsilon' \geq \varepsilon} \delta_{\max}(\varepsilon', P) = \delta_{\max}(\varepsilon, P)$$

by the monotonicity of $\delta_{\max}(., P)$. From this we easily obtain (18). ■

Proof of Theorem 3.9: To shorten notations we write

$$\begin{aligned} \mathcal{C}_{1,x}(f) &:= \mathcal{C}_{L_1,P(.|x),x}(f(x)) - \mathcal{C}^*_{L_1,P(.|x),x}\,, \qquad \text{and} \\ \mathcal{C}_{2,x}(f) &:= \mathcal{C}_{L_2,P(.|x),x}(f(x)) - \mathcal{C}^*_{L_2,P(.|x),x} \end{aligned}$$

for $x \in X$ and measurable $f : X \to \mathcal{T}$. Furthermore, for $s > 0$ we write

$$C(s) := \big\{x \in X : A(x) \neq \mathcal{T}, \text{ and } \delta_{\max}(h(x), P(.|x), x) \geq s\, h(x)\big\}\,.$$

By (9) and (20) we then obtain

$$\begin{aligned} \mathcal{R}_{L_1,P}(f) - \mathcal{R}^*_{L_1,P} &= \int_{B(s)} \mathbf{1}_A\big(x, f(x)\big) h(x)\, dP_X(x) + \int_{C(s)} \mathbf{1}_A\big(x, f(x)\big) h(x)\, dP_X(x) \\ &\leq \int \mathbf{1}_{B(s)} h\, dP_X + s^{-1} \int_{C(s)} \delta_{\max}\big(h(x), P(.|x), x\big)\, \mathbf{1}_A\big(x, f(x)\big)\, dP_X(x) \\ &\leq (c\,s)^\alpha + s^{-1} \int_{C(s)} \delta_{\max}\big(\mathcal{C}_{1,x}(f), P(.|x), x\big)\, dP_X(x) \\ &\leq (c\,s)^\alpha + s^{-1}\big(\mathcal{R}_{L_2,P}(f) - \mathcal{R}^*_{L_2,P}\big)\,. \end{aligned}$$

For $\alpha < \infty$, the assertion then follows by setting $s := (\alpha c^\alpha)^{-\frac{1}{\alpha+1}} \big(\mathcal{R}_{L_2,P}(f) - \mathcal{R}^*_{L_2,P}\big)^{\frac{1}{\alpha+1}}$ and using $\alpha^{-\frac{\alpha}{\alpha+1}} + \alpha^{\frac{1}{\alpha+1}} \leq 2$.
Furthermore, for $\alpha = \infty$ the assertion follows by setting $s^{-1} := 2c$. ■

Proof of Lemma 3.13: Let P be a distribution on $X \times Y$ with $P(.|x) \in \mathcal{Q}_{\min}(L)$ for all $x \in X$. We write $\bar{X} := X \times \mathbb{R}$ and $Z := \mathbb{R}$. Furthermore, for $\bar{x} = (x, t) \in \bar{X}$ and $t' \in Z$ we define

$$\begin{aligned} h(\bar{x}, t') &:= \mathcal{C}_{L,P(.|x)}(t') - \mathcal{C}^*_{L,P(.|x)}\,, \\ F(\bar{x}) &:= \{t' \in \mathbb{R} : h(\bar{x}, t') = 0\}\,, \qquad \text{and} \\ \varphi(\bar{x}, t') &:= |t - t'|\,. \end{aligned}$$

For the P-instance $\breve{L}_P$ of $\breve{L}$ we then have

$$\breve{L}_P(x, t) = \inf_{t' \in \mathcal{M}_{L,P(.|x)}(0^+)} |t - t'| = \inf_{t' \in F(\bar{x})} \varphi(\bar{x}, t')\,,$$

and consequently we obtain the assertion by part *iii)* of Lemma 5.1. ■

Proof of Theorem 3.16: For a fixed $\rho > 0$ we write $A_\rho = \{(Q,t) \in \mathcal{Q} \times \mathbb{R} : \breve{L}(Q,t) \geq \rho\}$. By Lemma 3.13 we then see that $\bar{L} := \mathbf{1}_{A_\rho}$ defines a template loss function, whose P-instance $\bar{L}_P$ is a detection loss with $h = \mathbf{1}_X$. Furthermore, we have $\mathcal{M}_{\bar{L},Q}(\varepsilon) = \{t \in \mathbb{R} : \breve{L}(Q,t) < \rho\} = \mathcal{M}_{\breve{L},Q}(\rho)$ for all $\varepsilon > 0$ and $Q \in \mathcal{Q}$, and thus we find

$$\delta_{\max,\bar{L},L}(\varepsilon, Q) = \delta_{\max,\breve{L},L}(\rho, Q) \;>\; 0\,.$$

In other words, L is $\bar{L}$-calibrated with respect to $\mathcal{Q}$. For $\varepsilon > 0$ Theorem 3.3 thus gives a $\delta > 0$ such that for $f : X \to \mathbb{R}$ with $\mathcal{R}_{L,P}(f) < \mathcal{R}^*_{L,P} + \delta$ we have

$$P_X\big(\{x \in X : \breve{L}_P\big(x, f(x)\big) \geq \rho\}\big) = \mathcal{R}_{\bar{L}_P,P}(f) - \mathcal{R}^*_{\bar{L}_P,P} < \varepsilon\,.$$

■

Proof of Proposition 3.19: We write $\alpha := \frac{pq}{p+1}$ and $\bar{L} := \breve{L}^\alpha$. Then $\bar{L}$ is a template loss function by Lemma 3.13, and since $\mathcal{M}_{\breve{L},Q}(\varepsilon) = \{t \in \mathbb{R} : \breve{L}(Q,t) < \varepsilon\}$ we easily find

$$\delta_{\max,\bar{L},L}(\varepsilon, Q) \;=\; \delta_{\max,\breve{L},L}(\varepsilon^{1/\alpha}, Q)\,, \qquad\qquad \varepsilon > 0,\, Q \in \mathcal{Q}.$$

Furthermore, we obviously have $\frac{q}{\alpha} \geq \frac{p+1}{p}$ and $\mathcal{R}^*_{\breve{L},P} = 0$, and therefore Theorem 2.18 yields

$$\|x \mapsto \breve{L}_P(x, f(x))\|^\alpha_\alpha \;=\; \mathcal{R}_{\bar{L}_P,P}(f) - \mathcal{R}^*_{\bar{L}_P,P} \;\leq\; \|b^{-1}\|_p^{\frac{\alpha}{q}} \Big(\mathcal{R}_{L,P}(f) - \mathcal{R}^*_{L,P}\Big)^{\frac{\alpha}{q}}\,.$$

■

Appendix

In this appendix we discuss some simple properties of convex functions. Recall that a function $f : I \to \mathbb{R}$ on an interval $I \subset \mathbb{R}$ is called *convex* if for all $x_1, x_2 \in I$ and all $\alpha \in [0,1]$ we have

$$f\big(\alpha x_1 + (1-\alpha)x_2\big) \;\leq\; \alpha f(x_1) + (1-\alpha) f(x_2)\,.$$

Furthermore, it is called *strictly convex* if this inequality is strict for all $x_1, x_2 \in I$ with $x_1 \neq x_2$, and all $\alpha \in (0,1)$. Obviously, every strictly convex function is also convex. Furthermore, it is well-known that convex functions $f : I \to \mathbb{R}$ are continuous on the interior of I. The following lemma describes some less trivial relations between the different notions of convexity:

Lemma A.1 *Let $f : I \to \mathbb{R}$ be a function. Then we have:*

i) If f is convex and satisfies $f(\alpha_0 x_1 + (1-\alpha_0)x_2) = \alpha_0 f(x_1) + (1-\alpha_0) f(x_2)$ for some $x_1, x_2 \in I$, $\alpha_0 \in [0,1]$, then actually for all $\alpha \in [0,1]$ we have

$$f\big(\alpha x_1 + (1-\alpha)x_2\big) \;=\; \alpha f(x_1) + (1-\alpha) f(x_2)\,. \tag{67}$$

ii) If f is continuous then f is convex if and only if for all $x_1, x_2 \in I$ we have

$$f\Big(\frac{x_1 + x_2}{2}\Big) \leq \frac{1}{2} f(x_1) + \frac{1}{2} f(x_2)\,. \tag{68}$$

iii) If f is continuous then f is strictly convex if and only if for all $x_1, x_2 \in I$ with $x_1 \neq x_2$ we have

$$f\Big(\frac{x_1+x_2}{2}\Big) < \frac{1}{2}f(x_1) + \frac{1}{2}f(x_2)\,. \tag{69}$$

iv) If f is uniformly convex and continuous then it is strictly convex. Conversely, if I is compact and f is strictly convex and continuous then it is actually uniformly convex.

Proof: *i).* This assertion can be shown using elementary calculations.
ii). This follows from [2, Thm. 8 and 10].
iii). If (69) holds then we have already seen that f is convex. Consequently, if f was not strictly convex we would have (67). However, by *i)* we could then assume $\alpha_0 = \frac{1}{2}$ which would give a contradiction.
iv) The first assertion follows from *iii)* and the second assertion is trivial. ■

Our next aim is to investigate the modulus of convexity. Although this concept, in an equivalent formulation, has already been introduced in 1966 (see [23], [17]) almost nothing that is useful for us, seems to be known (see however [5, 35] and the references therein for some general information on the modulus). Therefore, we present the following two lemmas which provide some ways to simplify the computation of $\delta_f(\varepsilon)$:

Lemma A.2 *Let $\emptyset \neq I \subset \mathbb{R}$ be an interval and $f : I \to \mathbb{R}$ be strictly convex. Then we have*

$$\delta_f(2\varepsilon) = \inf\left\{\frac{f(x-\varepsilon)+f(x+\varepsilon)}{2} - f(x) : x \text{ satisfies } x-\varepsilon \in I \text{ and } x+\varepsilon \in I\right\}, \qquad \varepsilon > 0.$$

Proof: For fixed $x_1 \in I$ we define $h_{x_1} : I \to [0,\infty)$ by $h_{x_1}(x_2) := \frac{f(x_1)+f(x_2)}{2} - f(\frac{x_1+x_2}{2})$, $x_2 \in I$. The fundamental theorem of calculus for Lebesgue integrals then shows that the derivative $h'_{x_1}(x_2)$ exists for almost all x_2, and an easy calculation shows $h'_{x_1}(x_2) = \frac{f'(x_2)}{2} - \frac{1}{2}f'(\frac{x_1+x_2}{2})$ for such x_2. Furthermore, since f is strictly convex and h_{x_1} has a unique minimum at x_1, we obtain $h'_{x_1}(x_2) < 0$ if $x_2 < x_1$, and $h'_{x_1}(x_2) > 0$ if $x_2 > x_1$. The fundamental theorem of calculus for Lebesgue integrals then shows that h_{x_1} is strictly decreasing on $(-\infty, x_1) \cap I$, and strictly increasing on $(x_1, \infty) \cap I$, and thus we have

$$\delta_f(2\varepsilon) = \min\Big\{\inf_{\substack{x_1 \in I \\ x_1 - 2\varepsilon \in I}} h_{x_1}(x_1 - 2\varepsilon), \inf_{\substack{x_1 \in I \\ x_1 + 2\varepsilon \in I}} h_{x_1}(x_1 + 2\varepsilon)\Big\} = \inf_{\substack{x_1+\varepsilon \in I \\ x_1-\varepsilon \in I}} h_{x_1-\varepsilon}(x_1+\varepsilon)\,,$$

where in the last step we used $h_{x_1-\varepsilon}(x_1+\varepsilon) = h_{x_1+\varepsilon}(x_1-\varepsilon)$. ■

With the help of the following lemma we can often estimate the modulus of convexity.

Lemma A.3 *Let $I \subset \mathbb{R}$ be a symmetric interval, i.e. $x \in I$ implies $-x \in I$. Then for all strictly convex, symmetric $f : I \to [0,\infty)$ and all $\varepsilon > 0$ we have*

$$\delta_f(2\varepsilon) = \inf_{\substack{x \geq 0 \\ x+\varepsilon \in I}} \frac{f(x-\varepsilon)+f(x+\varepsilon)}{2} - f(x) = \frac{1}{2} \inf_{\substack{x \geq 0 \\ x+\varepsilon \in I}} \int_x^{x+\varepsilon} \big(f'(t) - f'(t-\varepsilon)\big)dt\,.$$

Furthermore, if $I = \mathbb{R}$, then for all $x \geq 12\varepsilon$ we have

$$f(x) \geq \frac{2\delta_f(2\varepsilon)x^2}{\varepsilon^2}\,.$$

Proof: The first equation follows from Lemma A.2 and the symmetry assumptions. Furthermore, by the fundamental theorem of calculus for Lebesgue integrals we obtain

$$f(x+\varepsilon) + f(x-\varepsilon) = 2f(x) + \int_x^{x+\varepsilon} \bigl(f'(t) - f'(t-\varepsilon)\bigr)dt\,, \tag{70}$$

and hence the second equation follows. Finally, in order to show the last assertion we first observe that f has a minimum at 0, and hence we have $f'(t) \geq 0$ for all $t \geq 0$ where the derivative exists. We write $b := 2\delta_f(2\varepsilon)$, and $x_n := 2\varepsilon n$ for $n \geq 1$. These definitions together with (70) yield real numbers $t_n \in [x_n, x_n + \varepsilon]$, $n \geq 1$, that satisfy

$$b \leq \int_{x_n}^{x_n+\varepsilon} \bigl(f'(t) - f'(t-\varepsilon)\bigr)dt \;\leq\; \varepsilon\bigl(f'(t_n) - f'(t_n - \varepsilon)\bigr)\,,$$

i.e. we obtain $f'(t_n) \geq f'(t_n - \varepsilon) + \frac{b}{\varepsilon}$ for all $n \geq 1$. Furthermore we have $t_n - \varepsilon \geq x_n - \varepsilon = x_{n-1} + \varepsilon \geq t_{n-1}$ and hence $f'(t_n - \varepsilon) \geq f'(t_{n-1})$, $n \geq 2$. By induction we thus find $f'(t_{n+1}) \geq f'(t_1) + \frac{bn}{\varepsilon} \geq \frac{bn}{\varepsilon}$ for all $n \geq 1$. Now let $t \geq 6\varepsilon$ such that $f'(t)$ exists. Then there is an $n \geq 3$ with $2\varepsilon n \leq t < 2\varepsilon(n+1)$, and hence we get

$$f'(t) \geq f'(t_{n-1}) \geq \frac{b(n-2)}{\varepsilon} > \frac{b(x-6\varepsilon)}{2\varepsilon^2}\,.$$

Consequently, for $x \geq 12\varepsilon$ the fundamental theorem of calculus for Lebesgue integrals gives

$$f(x) = f(6\varepsilon) + \int_{6\varepsilon}^{x} f'(t)dt \geq \frac{b}{2\varepsilon^2}\int_{6\varepsilon}^{x}(t-6\varepsilon)dt = \frac{b(x-6\varepsilon)^2}{4\varepsilon^2} \geq \frac{bx^2}{\varepsilon^2}\,.$$

■

Example A.4 We write $I := [-B, B]$, and for $0 < p < 2$ we define $\psi : I \to [0, \infty]$ by $\psi(t) := |t|^p$, $t \in I$. Then for all $\varepsilon \in [0, B]$ we have

$$\frac{p(p-1)}{2}B^{p-2}\varepsilon^2 \;\leq\; \delta_\psi(2\varepsilon) \;\leq\; \frac{p}{2(p-1)^2}B^{p-2}\varepsilon^2\,.$$

To see this let $t \in [0,1]$ and $a \in (0,1)$. Then we have $(1-0)^a = 1 - a0$ and $-a(1-t)^{a-1} \leq -a$, and thus we obtain $(1-t)^a \leq 1 - at$ by the fundamental theorem of calculus. For $s := 1/t$ we hence find $(1 - \frac{1}{s})^a \leq 1 - \frac{a}{s}$ which is equivalent to $(s-1)^a \leq s^a - as^{a-1}$. Since in addition $s^{a-1} \leq (s-1)^{a-1}$ implies $s^a - s^{a-1} \leq (s-1)^a$ we have

$$as^{a-1} \leq s^a - (s-1)^a \leq s^{a-1} \tag{71}$$

for all $0 < a < 1$ and all $s \geq 1$. Now an easy calculation shows

$$\psi'(t) - \psi'(t-\varepsilon) = p\begin{cases} t^{p-1} - (t-\varepsilon)^{p-1} & \text{if } t \geq \varepsilon \\ t^{p-1} + (\varepsilon - t)^{p-1} & \text{if } 0 \leq t \leq \varepsilon\,. \end{cases}$$

Furthermore, for $s := \frac{t}{\varepsilon} \geq 1$ we have $\psi'(t) - \psi'(t-\varepsilon) = \varepsilon^{p-1}\bigl(s^{p-1} - (s-1)^{p-1}\bigr)$, and thus our preliminary considerations give

$$(p-1)\varepsilon t^{p-2} = (p-1)\varepsilon^{p-1}s^{p-2} \leq \psi'(t) - \psi'(t-\varepsilon) \leq \varepsilon^{p-1}s^{p-2} = \varepsilon t^{p-2}\,.$$

From this we can easily prove the assertion using basic calculations.

Example A.5 We write $I := [-B, B]$ and define $\psi : I \to [0, \infty]$ by $\psi(t) := -\ln \frac{4e^t}{(1+e^t)^2}$, $t \in I$. Then for all $\varepsilon \in [0, B]$ we have

$$\frac{e^\varepsilon - 1}{2e^\varepsilon}\ln\frac{e^B + e^{2\varepsilon}}{e^B + e^\varepsilon} \;\leq\; \delta_\psi(2\varepsilon) \;\leq\; \frac{e^\varepsilon - 1}{e^\varepsilon}\ln\frac{e^B + e^{2\varepsilon}}{e^B + e^\varepsilon}\,.$$

Indeed, an easy calculation shows $\psi'(t) = \frac{e^t-1}{e^t+1}$ for all $t \in I$, and hence we obtain

$$\psi'(t) - \psi'(t-\varepsilon) = \frac{e^t-1}{e^t+1} - \frac{e^t-e^\varepsilon}{e^t+e^\varepsilon} = \frac{2e^t(e^\varepsilon-1)}{(e^t+1)(e^t+e^\varepsilon)}$$

for all $t \in [0,B]$, $\varepsilon \in [0,B]$. Consequently we have

$$\frac{e^\varepsilon-1}{e^t+e^\varepsilon} \;\leq\; \psi'(t) - \psi'(t-\varepsilon) \;\leq\; 2\frac{e^\varepsilon-1}{e^t+e^\varepsilon}$$

for all $t \in [0,B]$, $\varepsilon \in [0,B]$. Furthermore, for $\varepsilon > 0$ an easy calculation gives

$$\inf_{x\in[0,B-\varepsilon]} \int_x^{x+\varepsilon} \frac{e^\varepsilon-1}{e^t+e^\varepsilon} dt = \int_{B-\varepsilon}^{B} \frac{e^\varepsilon-1}{e^t+e^\varepsilon} dt = \frac{e^\varepsilon-1}{e^\varepsilon}\Big(t - \ln\big(e^t+e^\varepsilon\big)\Big)\Big|_{t=B-\varepsilon}^{B} = \frac{e^\varepsilon-1}{e^\varepsilon} \ln \frac{e^B+e^{2\varepsilon}}{e^B+e^\varepsilon}\,.$$

From this we easily find the assertion.

The following two lemmas establish important properties of the Fenchel-Legendre bi-conjugate operation **. We begin with

Lemma A.6 *Let $B > 0$ and $\delta : [0,B] \to [0,\infty)$ be a monotone increasing function with $\delta(0) = 0$ and $\delta(\varepsilon) > 0$ for all $\varepsilon \in (0,B]$. Then for all $\varepsilon \in (0,B]$ we have*

$$\delta^{**}(\varepsilon) > 0\,.$$

Proof: Let us assume that there exists an $0 < \varepsilon \leq B$ with $\delta^{**}(\varepsilon) = 0$. Then we have $(\varepsilon, 0) \in \operatorname{Epi}\delta^{**} = \overline{\operatorname{co}\operatorname{Epi}\delta}$, and hence there exists a sequence $(\varepsilon_n, y_n) \in \operatorname{co}\operatorname{Epi}\delta$ with $\varepsilon_n \to \varepsilon$ and $y_n \to 0$. Furthermore, we have $\operatorname{co}\operatorname{Epi}\delta \subset \mathbb{R}^2$, and hence Carathéodory's theorem (see e.g. [24][p. 55]) guarantees that for all $n \geq 1$ there exist $\varepsilon_{n,1}, \varepsilon_{n,2}, \varepsilon_{n,3} \in [0,B]$, $y_{n,1}, y_{n,2}, y_{n,3} \in [0,\infty)$, and $\alpha_{n,1}, \alpha_{n,2}, \alpha_{n,3} \in [0,1]$ with

$$\begin{aligned}
\varepsilon_n &= \alpha_{n,1}\varepsilon_{n,1} + \alpha_{n,2}\varepsilon_{n,2} + \alpha_{n,3}\varepsilon_{n,3}\,,\\
y_n &= \alpha_{n,1}y_{n,1} + \alpha_{n,2}y_{n,2} + \alpha_{n,3}y_{n,3}\,,\\
1 &= \alpha_{n,1} + \alpha_{n,2} + \alpha_{n,3}\,,\\
y_{n,i} &\geq \delta(\varepsilon_{n,i})\,, \qquad\qquad i = 1,\ldots,3.
\end{aligned}$$

In addition we may assume $\varepsilon_{n,1} \leq \varepsilon_{n,2} \leq \varepsilon_{n,3}$ without loss of generality. Since this yields $\varepsilon_n = \alpha_{n,1}\varepsilon_{n,1} + \alpha_{n,2}\varepsilon_{n,2} + \alpha_{n,3}\varepsilon_{n,3} \leq \varepsilon_{n,3}$ we find $y_{n,3} \geq \delta(\varepsilon_{n,3}) \geq \delta(\varepsilon_n) \geq \delta(\frac{\varepsilon}{2}) > 0$ for large n. Recalling $y_n \to 0$ we thus obtain $\alpha_{n,3} \to 0$ which implies both $\alpha_{n,1} + \alpha_{n,2} \to 1$ and $\alpha_{n,1}\varepsilon_{n,1} + \alpha_{n,2}\varepsilon_{n,2} \to \varepsilon$. However, the latter convergence gives $\frac{\varepsilon}{2} \leq \alpha_{n,1}\varepsilon_{n,1} + \alpha_{n,2}\varepsilon_{n,2} \leq (\alpha_{n,1} + \alpha_{n,2})\varepsilon_{n,2}$ for large n, and hence we have $\varepsilon_{n,2} \geq \frac{\varepsilon}{4}$ for large n. Again this shows $y_{n,2} \geq \delta(\varepsilon_{n,2}) \geq \delta(\frac{\varepsilon}{4}) > 0$ for large n, and thus we find $\alpha_{n,2} \to 0$. Obviously, this yields both $\alpha_{n,1} \to 1$ and $\alpha_{n,1}\varepsilon_{n,1} \to \varepsilon$, and hence we obtain $\varepsilon_{n,1} \geq \frac{\varepsilon}{4}$ for large n. Finally, this gives $y_{n,1} \geq \delta(\varepsilon_{n,1}) \geq \delta(\frac{\varepsilon}{4}) > 0$ for large n and therefore we find $\alpha_{n,1} \to 0$ which contradicts the already found convergence $\alpha_{n,1} \to 1$. ■

Lemma A.7 *Let $B > 0$, and $\delta : [0,B] \to [0,\infty)$ be a continuous function with $\delta(0) = 0$. We define $\tilde\delta : [0,B] \to [0,\infty)$ by $\tilde\delta(\varepsilon) := \inf_{\varepsilon' \geq \varepsilon} \delta(\varepsilon)$, $\varepsilon \in [0,B]$. Then $\tilde\delta$ is monotonously increasing, and for all $\varepsilon \in [0,B]$ we have*

$$\delta^{**}(\varepsilon) = \tilde\delta^{**}(\varepsilon)\,.$$

In addition, if $\delta(\varepsilon) > 0$ for all $\varepsilon \in (0,B]$, then $\tilde\delta(\varepsilon) > 0$ for all $\varepsilon \in (0,B]$.

Proof: The first assertion is trivial and the third assertion directly follows from the continuity of δ. Therefore, it remains to show

$$\operatorname{co}\operatorname{Epi}\delta = \operatorname{co}\operatorname{Epi}\tilde{\delta} \tag{72}$$

since this equation immediately yields $\delta^{**} = \tilde{\delta}^{**}$. To establish (72) we first observe that $\tilde{\delta}(\varepsilon) \leq \delta(\varepsilon)$ for all $\varepsilon \in [0, B]$ and hence we have $\operatorname{co}\operatorname{Epi}\delta \subset \operatorname{co}\operatorname{Epi}\tilde{\delta}$. To prove the converse inclusion observe that it suffices to show $(\varepsilon, \tilde{\delta}(\varepsilon)) \in \operatorname{co}\operatorname{Epi}\delta$ for all $\varepsilon \in [0, B]$. Furthermore, we have $\tilde{\delta}(0) = 0 = \delta(0)$ and $\tilde{\delta}(B) = \delta(B)$ and hence we can restrict our considerations to pairs $(\varepsilon, \delta(\varepsilon))$ for $\varepsilon \in (0, B)$. Therefore let us fix an $\varepsilon \in (0, B)$. By the definition of $\tilde{\delta}$ we then find an $\varepsilon_+ \in [\varepsilon, B]$ with $\delta(\varepsilon_+) = \tilde{\delta}(\varepsilon)$. Furthermore, we have $\delta(0) \leq \tilde{\delta}(\varepsilon) \leq \delta(\varepsilon)$ and hence the intermediate value theorem applied to the continuous function δ gives us an $\varepsilon_- \in [0, \varepsilon]$ with $\delta(\varepsilon_-) = \tilde{\delta}(\varepsilon)$. Now, there exists an $\alpha \in [0, 1]$ with $\varepsilon = \alpha\varepsilon_+ + (1 - \alpha)\varepsilon_-$ and since our previous considerations showed

$$\tilde{\delta}(\varepsilon) = \alpha\tilde{\delta}(\varepsilon) + (1 - \alpha)\tilde{\delta}(\varepsilon) = \alpha\delta(\varepsilon_+) + (1 - \alpha)\delta(\varepsilon_-)$$

we obtain $(\varepsilon, \tilde{\delta}(\varepsilon)) \in \operatorname{co}\operatorname{Epi}\delta$. ■

References

[1] P.L. Bartlett, M.I. Jordan, and J.D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101:138–156, 2006.

[2] F.A. Behringer. Convexity is equivalent to midpoint convexity combined with strict quasiconvexity. *Optimization*, 24:219–228, 1992.

[3] C. Bennett and R. Sharpley. *Interpolation of Operators*. Academic Press, Boston, 1988.

[4] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Ann. Statist.*, submitted, 2004.

[5] D. Butnariu and A.N. Iusem. *Totally Convex Functions for Fixed Points Computation and Infinite Dimensional Optimization*. Kluwer, Dordrecht, 2000.

[6] A. Caponnetto. A note on the role of squared loss in regression. Technical report, Massachusetts Institute of Technology, 2005. `http://cbcl.mit.edu/projects/cbcl/publications/ps/caponnetto-squareloss-6-05.pdf`.

[7] C. Castaing and M. Valadier. *Convex Analysis and Measurable Multifunctions*. Springer-Verlag, Berlin, 1977.

[8] A. Christmann and I. Steinwart. Consistency and robustness of kernel based regression. *Bernoulli*, submitted, 2005. `http://www.c3.lanl.gov/~ingo/publications/ann-04b.pdf`.

[9] O. Dekel, S. Shalev-Shwartz, and Y. Singer. Smooth ε-insensitive regression by loss symmetrization. *J. Mach. Learn. Res.*, 6:711–741, 2005.

[10] C. Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, pages 973–978, 2001.

[11] L.M. Graves. *The Theory of Functions of Real Variables*. McGraw-Hill, New York, 1956.

[12] J.A. Hartigan. Estimation of a convex density contour in 2 dimensions. *J. Amer. Statist. Assoc.*, 82:267–270, 1987.

[13] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[14] K-U. Höffgen and Hans-U. Simon. Robust trainability of single neurons. In *Proceedings of the Computational Learning Theory (COLT) Conference*, pages 428–438, 1992.

[15] P.J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101, 1964.

[16] H. Kestelman. *Modern Theories of Integration*. Dover, New York, 1960.

[17] E.S. Levitin and B.T. Polyak. Convergence of minimizing sequences in conditional extremum problems. *Soviet Math. Dokl.*, 7:764–767, 1966.

[18] Y. Lin. Support vector machines and the Bayes rule in classification. *Data Mining and Knowledge Discovery*, 6:259–275, 2002.

[19] Y. Lin. A note on margin-based loss functions in classification. *Statist. Probab. Lett.*, 68:73–82, 2004.

[20] Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. *Machine Learning*, 46:191–202, 2002.

[21] D.W. Müller and G. Sawitzki. Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.*, 86:738–746, 1991.

[22] W. Polonik. Measuring mass concentrations and estimating density contour clusters—an excess mass aproach. *Ann. Statist.*, 23:855–881, 1995.

[23] B.T. Polyak. Existence theorems and convergence of minimizing sequences for extremal problems with constraints. *Soviet Math. Dokl.*, 7:72–75, 1966.

[24] R.T. Rockafellar and R.J.-B. Wets. *Variational Analysis*. Springer, 1998.

[25] A. Smola, B. Schölkopf, and K.-R. Müller. Convex cost functions for support vector regression. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of the International Conference on Artificial Neural Networks*, Perspectives in Neural Computing, pages 99–104, Berlin, 1998. Springer.

[26] I. Steinwart. Sparseness of support vector machines. *J. Mach. Learn. Res.*, 4:1071–1105, 2003.

[27] I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *IEEE Trans. Inform. Theory*, 51:128–142, 2005.

[28] I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *J. Mach. Learn. Res.*, 6:211–232, 2005.

[29] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, accepted, 2006. `http://www.c3.lanl.gov/~ingo/publications/ann-04a.pdf`.

[30] A. Tewari and P.L. Bartlett. On the consistency of multiclass classification methods. In *Proceedings of the 18th Annual Conference on Learning Theory, COLT 2005*, pages 143–157. Springer, 2005.

[31] A.B. Tsybakov. On nonparametric estimation of density level sets. *Ann. Statist.*, 25:948–969, 1997.

[32] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[33] V.N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[34] C. Zălinescu. On uniformly convex functions. *J. Math. Anal. Appl.*, 95:344–374, 1983.

[35] C. Zălinescu. *Convex Analysis in General Vector Spaces*. World Scientific, New Jersey, 2002.

[36] T. Zhang. Statistical analysis of some multi-category large margin classification methods. *J. Mach. Learn. Res.*, 5:1225–1251, 2004.

[37] T. Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32:56–134, 2004.